PB85-124139

Adjustment of Incidence Rates for Migration in Indirect Ecologic Studies

California Univ., Berkeley

Prepared for

Health Effects Research Lab.
Research Triangle Park, NC

Nov 84

U.S. Department of Commerce
National Technical Information Service
NTIS

PB85-124139

EPA-600/1-84-017
November 1984

ADJUSTMENT OF INCIDENCE RATES FOR MIGRATION IN INDIRECT ECOLOGIC STUDIES

by

Chin Long Chiang
Paul M. Conforti

School of Public Health
University of California
Berkeley, California 94720

Contract Numbers C1492NAEX and C2519NAET

Project Officer

Judy A. Stober
Toxicology and Microbiology Division
Health Effects Research Laboratory
Cincinnati, Ohio 45268

HEALTH EFFECTS RESEARCH LABORATORY
OFFICE OF RESEARCH AND DEVELOPMENT
U.S. ENVIRONMENTAL PROTECTION AGENCY
RESEARCH TRIANGLE PARK, NORTH CAROLINA 27711

REPRODUCED BY
NATIONAL TECHNICAL
INFORMATION SERVICE
U.S. DEPARTMENT OF COMMERCE
SPRINGFIELD, VA. 22161

TECHNICAL REPORT DATA (Please read Instructions on the reverse before completing)		
1. REPORT NO. EPA-600/1-84-017	2.	3. RECIPIENT'S ACCESSION NO. PB85 124139
4. TITLE AND SUBTITLE ADJUSTMENT OF INCIDENCE RATES FOR MIGRATION IN INDIRECT ECOLOGIC STUDIES		5. REPORT DATE November 1984 6. PERFORMING ORGANIZATION CODE
7. AUTHOR(S) Chin Long Chiang Paul M. Conforti		8. PERFORMING ORGANIZATION REPORT NO.
9. PERFORMING ORGANIZATION NAME AND ADDRESS Program in Biostatistics 140 Earl Warren Hall University of California Berkeley, CA 94720		10. PROGRAM ELEMENT NO. CBMCIA 11. CONTRACT/GRANT NO. C1492NAEX & C2519NAET
12. SPONSORING AGENCY NAME AND ADDRESS Health Effects Research Laboratory Office of Research and Development U.S. Environmental Protection Agency Research Triangle Park, NC 27711		13. TYPE OF REPORT AND PERIOD COVERED 14. SPONSORING AGENCY CODE EPA-600/11
15. SUPPLEMENTARY NOTES Judy Stober, Project Officer		

16. ABSTRACT

The overall objective of this research program was to develop a method for adjusting incidence rates for migration in studies relating environmental agents to diseases with long latent periods. Various methods of estimating migration and population change are considered.

An example of a situation necessitating this adjustment is described. Cancer incidence rates were compared for census tracts with varying levels of asbestos in drinking water. Since cancer has a long latent period, recent in-migrants would not have been exposed for sufficient periods of time to be at risk for cancer. Unless the in-migrants were equally distributed across census tracts, an analysis of the relationship between asbestos and cancer based on incidence rates would be biased.

This report reviews a number of measures of migration and population change as well as stochastic models of migration and stochastic models of population growth. Migration data available through the Census of Population and Housing of the Bureau of the Census are described. A method is developed that uses these data to estimate migration by census tract. This method is then applied to data from a U.S. Environmental Protection Agency supported project on the relationship between ingested asbestos and cancer and the results compared to the original analysis.

17. KEY WORDS AND DOCUMENT ANALYSIS			
a. DESCRIPTORS		b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field/Group
Asbestos Malignant Neoplasms Epidemiology Morbidity Contaminants Biostatistics Migrations	Statistical Analysis Stochastic Processes Demography	Indirect studies Ecologic studies Environmental studies	Biological and Medical Sciences Mathematical Sciences
18. DISTRIBUTION STATEMENT Release to Public		19. SECURITY CLASS (This Report) Unclassified 20. SECURITY CLASS (This page) Unclassified	21. NO. OF PAGES 80 22. PRICE

EPA Form 2220-1 (Rev. 4-77) PREVIOUS EDITION IS OBSOLETE

NOTICE

This document has been reviewed in accordance with U.S. Environmental Protection Agency policy and approved for publication. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

FOREWORD

The many benefits of our modern, developing, industrial society are accompanied by certain hazards. Careful assessment of the relative risk of existing and new man-made environmental hazards is necessary for the establishment of sound regulatory policy. These regulations serve to enhance the quality of our environment in order to promote the public health and welfare and the productive capacity of our Nation's population.

The complexities of environmental problems originate in the deep interdependent relationships between the various physical and biological segments of man's natural and social world. Solutions to these environmental problems require an integrated program of research and development using input from a number of disciplines. The Health Effects Research Laboratory, Research Triangle Park, NC and Cincinnati, OH, conducts a coordinated environmental health research program in toxicology, epidemiology, and clinical studies using human volunteer subjects. Wide ranges of pollutants known or suspected to cause health problems are studied. The research focuses on air pollutants, water pollutants, toxic substances, hazardous wastes, pesticides, and nonionizing radiation. The laboratory participates in the development and revision of air and water quality criteria and health assessment documents on pollutants for which regulatory actions are being considered. Direct support to the regulatory function of the Agency is provided in the form of expert testimony and preparation of affidavits as well as expert advice to the Administrator to assure the adequacy of environmental regulatory decisions involving the protection of the health and welfare of all U.S. inhabitants.

In this report, a method is developed to adjust disease incidence rates for migration in indirect studies of the relationship between environmental pollutants and diseases with long latent periods. The observational unit considered here is the census tract. Migration measures and models are reviewed as well as available census tract migration data.

F. Gordon Hueter, Ph.D.
Director
Health Effects Research Laboratory

ABSTRACT

The overall objective of this research program was to develop a method for adjusting incidence rates for migration in studies relating environmental agents to diseases with long latent periods. Various methods of estimating migration and population change are considered.

An example of a situation necessitating this adjustment is described. This was a study of the relationship between ingested asbestos and cancer. Cancer incidence rates were compared in census tracts with varying levels of asbestos in drinking water. Since cancer has a long latent period, recent in-migrants would not have been exposed for sufficient periods of time to be at risk for cancer. Unless the in-migrants were equally distributed across census tracts, an analysis of the relationship between asbestos and cancer based on incidence rates would be biased.

This report reviews a number of measures of migration and population change as well as stochastic models of migration and stochastic models of population growth. The stochastic models of migration include models that are time-independent and time-dependent. They vary in complexity from a simple in-migrant model to one in which in-, out-, and within-migration are included. The stochastic models of population growth extend this work to include birth and death considerations.

Migration data available through the Census of Population and Housing of the Bureau of the Census are described. A method is developed that uses these data to estimate migration by census tract. This method is applied to data from a U.S. Environmental Protection Agency supported project on the relationship between ingested asbestos and cancer. A reanalysis of this data with the addition of migration information is compared to the original results.

This report was submitted in fulfillment of contract Nos. C1492NAEX and C2519NAET by the University of California, Berkeley under the sponsorship of the U.S. Environmental Protection Agency. This report covers the period January 1, 1979, to September 30, 1981, and work was completed as of September 30, 1981.

TABLE OF CONTENTS

Foreword		iii
Abstract		iv
Tables		vi
1.	Introduction	1
2.	Measurement of Variables	4
3.	Description of Observed Cases and Populations in Terms of Migrant and Non-Migrant Sub-Groups	11
4.	Nonstochastic Migration Measures	16
5.	Stochastic Models of Migration	27
6.	Stochastic Modelling of Population Growth	41
7.	Census Migration Data	46
8.	Estimation of Migration by Census Migration Data	52
9.	Variables in Asbestos-Cancer Data Analysis	54
10.	Results	56
11.	Comparison of Original Asbestos-Cancer Study Results with Exposure Study Results	68
12.	Discussion	70
References		72

TABLES

Number		Page
1	Population of San Francisco Bay Area by County and Decade for 1950, 1960 and 1970	3
2	Cancer Sites Analyzed	55
3	Descriptive Statistics for Analyzed Variables	58
4	Pearson Product Moment Correlations and Significance for Observed to Expected Cancer Ratios and Asbestos by Site and Sex	61
5	Pearson Product Moment Correlations and Significance for Observed to Expected Cancer Ratios and Years of Exposure by Site and Sex	62
6	Regression Coefficients and Significance for Asbestos and Years of Exposure by Site and Sex	63
7	Regression Coefficients and Significance for Asbestos X Years of Exposure by Site and Sex	66
8	Significant Cancer Body Sites ($p<.01$) for Asbestos X Years of Exposure by Sex with Significance	67
9	Significant ($p<.01$) Regression Coefficients and Significance for Asbestos for Original and Reanalysis Studies by Sex	69

SECTION 1

INTRODUCTION

The relationship between inhaled asbestos fibers and cancers of the respiratory system in human populations is well documented. This knowledge prompted the hypothesis that asbestos fibers might be carcinogenic in digestive system tissue following exposure by ingestion of water containing asbestos and that asbestos fibers might also be carried from the digestive tract to other parts of the body such as the pancreas, liver, lymphatic system, etc. where cancer might occur.

The United States Environmental Protection Agency funded a study at the University of California, Berkeley, investigating the relationship between asbestos in domestic water supplies (tap or drinking water) and cancer of various sites of the body. An indirect design was used in this study. An indirect study is one in which the observational unit is an aggregate of subjects. Information is obtained for the aggregate rather than for individual subjects. In this case the observational units were geographical units called census tracts. In a direct study each subject's asbestos intake and cancer status would be measured. In the indirect study, asbestos content in water was measured for each census tract by one or more water samples and cancer was measured by the incidence rates (number of cases divided by mid-study period populations). Census tracts rather than individuals were compared.

The five-county San Francisco Bay Area was chosen for study since it yielded a natural experiment situation. Water supplies in the Bay Area vary in asbestos fiber content from 10^7 fibers per liter of water to virtually no fibers per liter (although the lowest detectable limit of asbestos fibers in water is approximately 10^4 fibers per liter). The asbestos in water occurs naturally due to the state rock -- serpentine, which contains asbestos and is prevalent in the coastal range of the mountains in the Bay Area. Some of the Bay Area's water reservoirs are constructed in these mountains with serpentine rock as their bases. Asbestos fibers are leeched from the rock into the water. Three of the five Bay Area counties are served with water from these reservoirs. The other two counties use water that is virtually free of asbestos fibers.

These five counties consist of 722 census tracts (geographical units used by the Census Bureau for enumeration of the population) according to the 1970 Census of Population. These census tracts were the observational units for the asbestos-cancer study. The Bay Area had a population of 3,109,519 in 1970. The census tracts averaged approximately 4,300 people each. The Census Bureau tries to construct census tracts of equal or

nearly equal population size. Data on covariables for census tracts such as median family income, percentage of population working in asbestos-related industries, percentage of population married, and mean years of schooling were obtained from the 1970 Census of Population.

The Tumor Registry of the California State Department of Health provided the cancer incidence data. These data included most major body site cancers. Of major interest were body site cancers of the thorax and abdomen. This information covered the three year period: 1969-1971. This time period was chosen since the 1970 Census centered in this period. Site specific cancer rates were formed from the cancer data and population data from the 1970 Census. A site specific cancer rate for a census tract for the three-year period is the number of occurrences of cancer of a given body site for that census tract for 1969-1971 divided by the population in that census tract in that time period. In essence, the cancer rates were correlated with asbestos values across census tracts. More specifically, the cancer rates were age-adjusted by the indirect method using aggregated five-county rates as the standard. The age adjusted rates were entered in a multiple regression as the dependent variable with the aforementioned covariables and asbestos values entered as the independent variables. The relationship between asbestos in water and cancer of various body sites then can be assessed by examination of regression coefficients (1).

The problems in the assessment of the relationship between asbestos and cancer in this situation are that: 1) the reported cancer occurrences are not linked temporally to asbestos, and/or 2) the population figures in a particular time period may not correctly represent the true population at risk. Both of these problems involve the long latent period (time between exposure to the carcinogenic agent and manifestation of cancer symptoms) coupled with the mobility of the population. These problems result in a distortion of the asbestos-cancer relationship that is observed by the introduction of bias in the age adjusted incidence rates. This distortion may occur in a number of ways. An in-migrant may develop cancer unrelated to asbestos exposure and be counted by the Tumor Registry. This would result in artificially inflated incidence rates as far as the asbestos-cancer relationship is concerned. An in-migrant might have arrived in time to have been counted in the 1970 census but not in time to have been exposed long enough (according to the latent period which in the case of cancer is estimated to be between 20 and 40 years) to be at risk of developing cancer. This would result in deflated incidence rates. An out-migrant that spent many years in the Bay Area and was counted in the census but did not spend the entire study period in the Bay Area might also deflate the incidence rate by manifesting cancer after the migration so that the Tumor Registry would miss the case. There is movement within the Bay Area that would also tend to bias the incidence rates. Migration from high to low asbestos census tracts or vice versa could bias the rates.

If migration could be assumed to be uniformly distributed across time and space then the problem of biased incidence rates would not exist. The relationship between asbestos and cancer would not be distorted in a relative sense but the magnitude of occurrence of cancer as expressed in each of the incidence rates would be inflated or deflated depending on the excess of out-

migrants or in-migrants, respectively. It is apparent that a large number of in-migrants have entered the Bay Area over the last several decades. The 1950 population of the five-county San Francisco Bay Area was 2.1 million compared with the 3.1 million of 1970. Furthermore this influx of people was not uniform across census tracts as evidenced by a breakdown of population by counties for 1950, 1960 and 1970 (2), (3), (4).

TABLE 1. POPULATION OF SAN FRANCISCO BAY AREA BY COUNTY AND DECADE FOR 1950, 1960 and 1970

County	1950	1960	1970
Alameda	740,315	908,209	1,073,184
Contra Costa	298,984	409,030	558,389
Marin	85,619	146,820	206,038
San Francisco	775,357	740,316	715,674
San Mateo	235,659	444,387	556,234
Total	2,135,934	2,648,762	3,109,519

The assumption of uniformly distributed in-migrants appears tenuous. Out-migrants, while not as large a group numerically as in-migrants also contribute to the bias by both changing the number of people observed in each tract as well as possibly altering the distribution of population over the census tracts. Within-migration exacerbates the problem by mingling the exposed and unexposed groups. Therefore, the observed relationship between asbestos and cancer would appear to be biased when movement of the population is not controlled for. Part of the problem is modelling the dynamics of population flow so that estimates of the actual population in each tract that is exposed or not exposed can be obtained in order to clarify the direction of the bias. The model must include in-migrants, out-migrants, and within-migrants.

The other part of the problem is to show how to use the data about at-risk and not-at-risk populations generated by the model to adjust the incidence rates in considering the relationship between asbestos and cancer. This would include the estimation of model parameters by currently available data or by information obtained from sampling. It would also include the analysis of adjusted incidence rates by multiple regression as outlined above.

SECTION 2

MEASUREMENT OF VARIABLES

The major independent variable was chrysotile asbestos fiber content of census tract drinking water. Water sampling was undertaken in order to determine the average number of chrysotile asbestos fibers per liter in drinking water in each census tract. The drinking water distribution systems of the five Bay Area counties were documented for current and historical supplies. This was necessary because of the need to obtain samples that would best typify the chrysotile asbestos exposure to the population via their drinking water over the last 40 years, the possible maximum latent period for asbestos-induced tumors. A summary description of the major sources of drinking water, treatment method, and distribution extant in the five counties follows:

1. San Francisco -- Water is imported from the Hetch Hetchy Reservoir, Yosemite National Park, which is located in the Sierra Nevada. The major impounding reservoirs for the Hetch Hetchy water are Crystal Springs and San Andreas Reservoirs, which are both located on the San Francisco Peninsula. The main historical supply to San Francisco has been Hetch Hetchy Reservoir water supplied through the Hetch Hetchy pipeline to the Crystal Springs Reservoir, with a portion of the Crystal Springs water being further drained into the San Andreas Reservoir. These waters were not filtered and were treated only for pH control and chlorinated for disinfection. Beginning in the early 1970's much of the Hetch Hetchy water was delivered directly to San Francisco via a bypass pipe around Crystal Springs Reservoir. Crystal Springs water is now delivered primarily into San Andreas Reservoir, where it is filtered before distribution to customers. There are some 25 different pressure zones in the San Francisco drinking water distribution system, which are formed throughout the city by closed valves between different parts of the pipe system. There are five transmission mains from the Peninsula reservoirs. These pipelines discharge into one or more of six receiving reservoirs throughout the city.

2. San Mateo -- The major source of water to this county is the Hetch Hetchy system. There are about 23 different water districts in the county that purchase water from the San Francisco City and County Water District. This study focused primarily on surface water supplies, excluding most well systems because of their limited area and problems of defining the distribution of these well systems over time. Thus, the major sources analyzed in this county were San Andreas, Crystal Springs, and San Francisco By-Pass water.

3. Marin -- Local runoff from the Coast Range provides the entire water supply for the Marin Municipal Water District, which presently services the major population areas of this county. Runoff is collected in four major reservoirs in the Mt. Tamalpais region and is eventually treated by Bon Tempe (constructed in 1959) or San Geronimo (constructed in 1962) treatment plants before distribution to consumers. Both treatment plans involve the usual water treatment procedures of flocculation, sedimentation, filtration, chlorination. During some extremely dry years, water has been imported from the Russian River to service the northern portion of the Marin Municipal Water District. Starting in June 1977, water was imported from the Sacramento-San Joaquin Delta via a pipeline from the Richmond pumping station in Contra Costa County to Marin, to provide water during the most extreme drought condition in 1977. These imported sources of water are anomalous situations and not representative of the water normally received by Marin residents over the last 40 years. Water samples in Marin County were chosen to represent normal runoff courses.

4. Contra Costa -- Approximately one-half of Contra Costa County received water from the Sacramento-San Joaquin Delta area. This is either directly from the Sacramento-San Joaquin River system or indirectly from the Contra Costa Canal, the intake of which is located in the Delta south of the San Joaquin River at Rock Slough. Most of this water is treated at the Bollman Treatment Plant and distributed by the Contra Costa County Water District. Two smaller capacity treatment plants service the cities of Antioch and Pittsburg. All three treatment plans use coagulation, flocculation, sedimentation, filtration, and chlorination. The other half of the county is serviced primarily by the East Bay Municipal Utility District (described under Alameda County).

5. Alameda -- Most of the water to Alameda County is obtained from the Mokelumne River and Pardee Reservoir, located in the Sierra Nevada north of Yosemite. The water from the source is distributed to various locations once it arrives in the Bay Area. Some of it goes directly to the Walnut Creek, Lafayette, and Orinda Filter Plans from the aqueduct. These plants use filtration, chlorination, and some lime addition. At times, flocculation has been used for turbidity control (reduction of muddiness or cloudiness). The Lafayette Reservoir also receives water directly from the aqueducts by pumping. Water for the Upper San Leandro Reservoir is pumped from Lafayette and is eventually treated before distribution. Water arriving at Orinda flows by gravity to San Pablo Reservoir, which feeds the Sobrante and San Pablo Treatment plants. Upper San Leandro, San Pablo, and Sobrante Treatment Plants employ aeration, coagulation, sedimentation, filtration, chlorination, and pH control.

After the process of documenting the area's water distribution system was completed, each census tract was coded as to which of the 74 water distribution companies served the tract and what type of water was the primary source of tract water. A sampling program was designed to determine the levels of asbestos within the distribution systems. Analyses were made on raw and finished water, and from samples taken from various sites within a given distribution system, the continuity of values throughout the distribution area was determined. Most samples were taken from household taps since

asbestos determination is needed to be made at the consumer use point in the distribution system. Stored emergency waters were analyzed also, some dating back to 1908. Although not many of these water caches were available, they yielded some historical background data on the presence of asbestos in Bay Area water.

The water samples were collected in one-half gallon glass bottles, coded for documentation, and then analyzed on a "blind" basis. The water samples were filtered through 0.45 micrometer size Millipore filters. The Millipore filter material, which is a cellulose ester, was removed by acetone vapor in a cold finger Soxhlet extractor. This is accomplished by first covering collodion-coated 200-mesh nickel grids with a thin layer of carbon in a high vacuum optical coater and then placing the grids on the cold finger. Discs, one-eighth inch in diameter, of the Millipore filter were placed, particulate side down, on the prepared grids. The cold finger was placed into the extractor cold and the acetone reservoir heated to a temperature sufficient to produce a very light tremor, i.e., just below the boiling point of acetone. The Millipore filter material was dissolved in 6-12 hours, leaving any particulate matter which was on the filter on the grid.

The grids were examined and counted using a Siemen Elmiskop electron microscope at magnification of 100,000 x optical magnification. One grid opening of the 200-mesh grid is considered as one field. In cases of samples heavily loaded with asbestos fibers, the number of fields necessary to get a total of 100 fibers were counted. In cases of samples with moderate to small amounts of fibers or with less than the level of the lowest detectable limit, 20 fields were counted. In other words, whichever came first of either 100 fibers or 20 fields were counted. The viewing screen on the microscope had fiducial marks used for sizing of the fibers such that at 10,000 x magnification, 2 mm on the screen represented 0.2 micrometers. The number of fibers counted were related to the area of a grid opening, the filter area, and the volume of water filtered from the original sample. The count was reported as the number of fibers per liter of sample. The lowest detectable limit of asbestos concentrations in finished water was between 1×10^4 to 1×10^5 fibers per liter depending on the degree of turbidity and consequent amount of water sample filtered.

The concentration of fibers per liter (f/l) is calculated by the formula:

$$f/l = \frac{\text{(average \# fibers/grid opening) x filter area x 1000 ml}}{\text{area of grid opening x sample volume in ml}}$$

In many of the water samples unidentified fibers were observed. Chrysotile asbestos fibers were identified and distinguished from other fibers by a combination of morphology and selected area diffraction patterns using the transmission electron microscope. The unidentified fibers could have been such things as amphibole asbestos fibers, fibrous talc, or halloysite (a type of fibrous clay).

In total, 353 water samples were analyzed. Asbestos values were assigned to census tracts by calculating the arithmetic mean of finished and tap water samples from a given water distribution area and assigning that value to all the tracts in that area. Two groups of asbestos values were assigned to the tracts. Where possible, each water district was subdivided by various sources, treatment processes, and/or pressure zones. "Fine" values were derived from taking the arithmetic mean of all finished and tap water chrysotile fiber counts for each of these subdivisions and were assigned to individual tracts. Sometimes this value was taken from a single asbestos count per tract (s); other times it was an average of all values within the subdivision. When a larger water purveyor supplied several small districts, each of the smaller districts' values were averaged separately for assigning these "Fine" values. Although this sometimes results in different values for the same basic source and supply system, this method did account for changing practices among the smaller water districts.

For the "Gross" set of values, all asbestos values for an individual district or system (whichever was larger) were averaged (arithmetic mean) and that value applied to all tracts served by that system. It was felt that this grouping would help mitigate some of the weaknesses of trying to characterize water quality with respect to asbestos fibers for 722 census tracts with only 353 grab samples.

The dependent variable for this study was the Standard Morbidity Ratio (SMR). This is the ratio of observed to expected (O/E) number of cancer cases. In order to obtain the necessary SMRs, cancer incidence case data and 1970 census population data were used in an indirect age-adjustment procedure using standard rates, which in this case were calculated rates for the whole area (SMSA). The standard rates were applied to the population distribution of each census tract to obtain the expected number of cases in each census tract if the rates of the standard population (SMSA) prevailed. This procedure, which is described in further detail in this section, adjusts for age, that is, eliminates any age distribution differences between census tracts. If the resulting SMR (O/E) is less than 1, fewer cases were observed in that tract than would be expected based on the rates for the entire area. If the SMR is greater than 1, more cases were observed than expected.

The cancer incidence case data for 1969-71 was obtained from the San Francisco Bay Area Resource for Cancer Epidemiology (RCE) which is located in the State of California Department of Health Services. These data were collected as part of the Third National Cancer Survey (TNCS). The Third National Cancer Survey collected medical and other data on all incidence (newly-diagnosed) malignancies during the three-year period 1969-71 for nine areas of the U.S. The areas included two entire states, Iowa and Colorado, and seven metropolitan areas, including the San Francisco-Oakland Standard Metropolitan Statistical Area. The design and operations of the TNCS are detailed in a 1975 Third National Cancer Survey Monograph.

The tumor site and type were coded according to the 1968 Manual of Tumor Nomenclature and Coding. A noteworthy feature of the TNCS was the classification by morphologic type of tumor (sometimes called histologic type). This

classification defined over 200 pathologically different cell types of malignancy.

The 1969-71 three-year period was chosen specifically because the 1970 Census of Population would provide accurate estimates of the populations at risk and data on other attributes of those populations. Thus the need for intercensal population estimates was eliminated. Accurate incidence rate calculations resulted from the Third National Cancer Survey because all newly diagnosed cases for an area were collected, and with use of 1970 Census population data, accurate incidence rates were calculable.

The cancer incidence data included most major body sites. For each case there were codes for age, sex, race, body site of the malignancy, histologic cell type, diagnosis data, hospital, method of diagnosis, and census tract. The major sites excluded from analysis where skin and bone tumors.

For each census tract counts of cases were made for each tumor body site, which was further subdivided into separate records for each sex-race group, and each of those records was divided into five-year age groups. The races were white, nonwhite, and total. This procedure of counting cases by age, sex, race, and site specific cells for each census tract resulted in the observed numbers of cases, needed for the numerator in the Standard Morbidity Ratio.

In addition to individual cancer sites, some logical groupings of sites were analyzed. An example is the digestive tract which includes sites from the esophagus to the rectum. The sites analyzed were identified by their Resource for Cancer Epidemiology (RCE) and 8th Revision International Classification of Disease (ICD) codes.

Population data for use with the three-year 1969-71 cancer incidence data was taken from the 1970 Census of Population. The mid-point of the three-year period was June 30, 1970, which was very close to the data of the 1970 Census -- April 1, 1970. Therefore, the 1970 Census population data was used directly for SMSA and census tract incidence rates.

The Second Count Summary Tape, File A, Table 2 of the 1970 Census of Population was used for obtaining race specific (white, nonwhite, and total), sex specific (male and female), age specific (five-year age groups -- 0-4, 5-9, ..., 70-74, 75+) population counts for each tract and for the SMSA. Table 2 contains age, sex, race specific population numbers for the SMSA. Tables 3 and 4 show SMSA observed cases and SMSA incidence rates (one-year), respectively.

The age adjustment procedure was the indirect method. Observed numbers of cases for a site were compiled for each age group in each tract for each sex-race group, and expected numbers were calculated by multiplying the SMSA age, race, sex, site specific rates as standard rates to the age, sex, race population distribution of each census tract.

These procedures yielded observed and expected numbers of cancer cases in each census tract, for each body site, for each sex-race group, by

five-year age groups. The ratio of the observed-to-expected numbers of cases for an age group or any combination of age groups, is the Standard Morbidity Ratio (SMR), which was used as the dependent variable in the analysis.

Other variables were organized by census tract for inclusion in the analysis. These variables were chosen because of their possible influence in confounding the relation between asbestos levels in drinking water and cancer incidence.

Socioeconomic status is well known to affect cancer rates. For instance, stomach and lung cancers appear to be associated with low, and colon cancer with high socioeconomic status. Median family income and median school years completed were used as indicators of socioeconomic status.

Since married people are known to have lower disease rates than those single, separated, widowed, or divorced, marital status data for each tract was utilized. The percentages of the tract population 14 years old and over who were single, separated, widowed, or divorced in 1970 (all except married) were added together in a variable called marital status.

The variable to estimate the numbers of workers in each tract who were employed in industries that might have high exposures to asbestos were used in a variable named Industry. The number of construction workers was added to the number of electrical workers and textile workers and then divided by the total population.

Cancer rates among many migrant population groups from other countries differ from those of the native U.S. population. Countries for which migrant rates exceed those of the U.S. native population were included in the analysis. Census data on native and foreign-born persons of foreign stock was used in a variable called Country of Origin. The number from each country was divided by the total population of each census tract forming a ratio between 0 and 1. When two or more countries were included for a site, the ratios of those countries were added forming a single index amount of foreign stock from the specific countries in question.

Data on residence stability and year moved into unit were also organized by census tract. They were both made into ratios that would serve as indices of stability for the tracts.

There appears to be a synergistic effect between asbestos exposure and cigarette smoking in causing lung cancer in asbestos workers. In addition, cigarette smoking is implicated in varying degrees in the causation of cancer of many other body sites. Consequently, a variable reflecting cigarette smoking was included in the analysis. Since census tract cigarette consumption data for the study area was unavailable, rates of squamous cell bronchogenic carcinoma, the type of lung cancer that is most related to cigarette smoking, was used as a surrogate parameter for cigarette consumption. The Standard Morbidity Ratio for each sex-race group for squamous cell bronchogenic carcinoma was assigned to each corresponding population group for each census tract. The assumption here in trying to use rates of a specific type

of lung cancer as a surrogate parameter for cigarette smoking was that there is a geographic correlation between smoking and lung cancer rates.

The variable titled Occupation was used as another indication of asbestos exposure at the work place. Craftsmen and kindred workers were added to laborers, except farmers and divided by the total employed population, 16 years old and over in each census tract (1).

SECTION 3

DESCRIPTION OF OBSERVED CASES AND POPULATIONS IN TERMS OF MIGRANT AND NON-MIGRANT SUB-GROUPS

As mentioned earlier, the analyses to test for the relationship between ingested asbestos and cancers of various body sites used the indirect age-adjusted rates and standard morbidity ratio as dependent variables. The following analysis applies also if elements of the standard morbidity ratio (the observed and expected cases of cancer) are used separately such as in a regression procedure in which the observed cases of cancer are treated as a dependent variable and the expected cases are inserted in the regression equation as an independent variable.

The indirect age-adjusted incidence rate took the form:

$$\text{I.A.-A.R.}_{ik} = \frac{O_{ik}/P_i}{\sum_{j=1}^{n} P_{ji} \frac{O_{jk,Standard}}{P_{j,Standard}} / P_i} \cdot \frac{O_{k,Standard}}{P_{Standard}},$$

where I.A.-A.R._{ik} is the indirect age-adjusted incidence rate in the i-th area for the k-th body site (in this case i represents the census tracts or the area outside of the studied SMSA, i=1,...,723; and k represents the body site grouping of cancer studied, k=1,...,35); O_{ik} is the observed cases of cancer k in area i, P_i is the population in area i; O_{ik}/P_i is the observed incidence rate for cancer k in area i; P_{ji} is the population in area i and age group j (15 five-year age groups for 0-74 years of age and an age group for persons 75 years of age and up, j=1,...,16); $O_{k,Standard}$ is the observed cases in standard population (the Standard Metropolitan Statistical Area - SMSA which is the sum of 722 census tracts); $P_{Standard}$ is the population in the standard area; $O_{k,Standard}/P_{Standard}$ is the observed incidence rate in the standard population, $O_{jk,Standard}$ is the observed cases of cancer k in the j-th age group in the standard population; $P_{j,Standard}$ is the population in the j-th age group in the standard population; $O_{jk,Standard}/P_{j,Standard}$ is the age-specific incidence ratio for age group j in the standard population for the k-th cancer; $\sum_{i=1}^{n} P_{ji} \frac{O_{jk,Standard}}{P_{j,Standard}} / P_i$ is the expected incidence

rate in area i for the k-th cancer if the age specific rates for the standard population are applied to the population distribution in area i; $\sum_{j=1}^{n} P_{ji} \frac{O_{jk,Standard}}{P_{j,Standard}}$ is the expected cases of cancer k in area i.

The other variables, the standard morbidity ratio, took the form:

$$S.M.R._{ik} = \frac{O_{ik}/P_i}{\sum_{j=1}^{n} P_{ji} \frac{O_{jk,Standard}}{P_{j,Standard}} / P_i} = \frac{O_{ik}}{\sum_{i=1}^{n} P_{ji} \frac{O_{jk,Standard}}{P_{j,Standard}}} = \frac{O_{ik}}{E_{ik}} .$$

Obviously, whether the analysis uses the indirect age-adjusted rate, the standard morbidity ratio, the observed incidence rate, or the observed and expected cases separately, the problem of migration will enter the analysis through one or more of O_{ik}, P_i, $O_{jk,Standard}$, $P_{j,Standard}$, P_{ji}. The age groups were five-year categories from 0-74 years of age plus a category for persons 75 years of age and up yielding 16 groups. The subscript j represents the 16 age groups and takes the values j=1-16. The standard population in the asbestos-cancer study was the sum of the populations across all census tracts in the five-county area (the Standard Metropolitan Statistical Area).

If the assumption of a 20 year latent period is applied, then O_{ik}, P_i, $O_{jk,SMSA}$, and $P_{j,SMSA}$ need adjustment. In O_{ik} and $O_{jk,SMSA}$, the observed cases of cancer in individuals less than 20 years of age should be removed. If $O_{ik} = \sum_{j=1}^{16} O_{jik}$ is all of the observed cases of cancer k in the i-th census tract, then the observed cases in the i-th census tract after the removal of cases in people less than 20 years of age is $O'_{ik} = \sum_{j=5}^{16} O_{jik}$ since j=1 represents ages 0-4, j=2 represents ages 5-10, j=3 represents ages 10-14, and j=4 represents ages 15-19. Analogously, the observed cases of cancer k for the SMSA are $O_{k,SMSA} = \sum_{j=1}^{16} O_{jk,SMSA}$ and the number of cases after removal of the cases of people under the age of the latent period are $O'_{k,SMSA} = \sum_{j=5}^{16} O_{jk,SMSA}$.

For population figures the same reasoning under the assumption of a twenty year latent period can be used to reduce $P_i = \sum_{j=1}^{16} P_{ji}$ to

$$P'_i = \sum_{j=5}^{16} P_{ji} \text{ and } P_{SMSA} = \sum_{j=1}^{16} P_{j,SMSA} \text{ to } P'_{SMSA} = \sum_{j=5}^{16} P_{j,SMSA}.$$

The results of these adjustments yield the variables:

$$I.A.\text{-}A.R.'_{ik} = \frac{O'_{ik}/P'_i}{\sum_{j=5}^{16} P_{ji} \frac{O_{jk,SMSA}}{P_{j,SMSA}} / P'_i} \cdot \frac{O'_{k,SMSA}}{P'_{SMSA}},$$

$$I.R.'_{ik} = \frac{O'_{ik}}{P'_i},$$

$$S.M.R.'_{ik} = \frac{O'_{ik}/P'_i}{\sum_{j=5}^{16} P_{ji} \frac{O_{jk,SMSA}}{P_{j,SMSA}} / P'_i} = \frac{O'_{ik}}{\sum_{j=5}^{16} P_{ji} \frac{O_{jk,SMSA}}{P_{j,SMSA}}}.$$

The reasoning behind the removal of cases and population as done above is that since a latent period of 20 years is assumed, the cases of cancer observed in people less than 20 years of age had to be caused by or related to some factor(s) other than asbestos and the population under 20 years of age is not yet considered to be at risk of developing cancer.

Having eliminated the nonlatent period people from the sample of population and cases we turn our attention to the migrants. Certainly O'_{ik} contains cases that are not related to asbestos because of a combination of latent period and migration. In each of O_{jik} for j=5-16, there are the cases of individuals who had lived in the i-th census tract (i=1,...,722 actual census tracts and not i=723 which represents the outside of SMSA region) for less than 20 years and the complementary group of cases of people who had lived in the i-th census tract for 20 or more years. In the fifth age group (20-24 years of age), $O_{5ik} = O_{5ik}$(lived 20-24 years in i) + O_{5ik}(lived <20 years in i). This is perhaps the crudest view of migration since the cases of people who had lived in the i-th census tract for less than 20 years may have come from outside the study area or from inside the study area (from one or more other census tracts with the same and/or different levels of asbestos exposure). Cases of people who had lived in the i-th census tract for less than 20 years, from census tract(s) with the same asbestos content in the water supply such that their total exposure at that particular level of asbestos for $\geq$20 years could then be included with the O_{5ik}(lived 20-24 years in i) to yield the number of observed cases of cancer k that are possibly related to asbestos exposure.

Another assumption, besides that concerning the latent period, needs to be made concerning the intensity of exposure to asbestos fibers. The question raised here refers to a comparison in effect of different levels of exposure over different lengths of time. It needs to be decided whether or not 10 years of exposure at, say 1×10^6 fibers per year is equivalent in effect to exposure of 2×10^6 fibers per liter for five years. It is clear that assumptions about the latent period enter into these considerations. For instance, neither of the above two hypothetical situations would yield a cancer case that would be attributed to asbestos if an assumption of a 20 year latent period was operating. Under the 20 year latent period assumption, if one individual was exposed to water with 2×10^6 fibers per liter for 20 years and another was exposed to water with 1×10^6 fibers per liter for 40 years, the question would arise as to the equivalence of exposure. If the threshold level (amount necessary to cause disease) is based on the daily level of exposure rather than a total or cumulative amount of exposure, then a distinction would need to be made between the above two cases of exposure. If the threshold level is a composite of latent period and cumulative amount of exposure, then the above two cases of exposure would be considered equivalent.

A more refined partitioning of each O_{jik} is needed to enumerate all of the possible situations from which the cancer cases were obtained. For the purposes of this partitioning the subscripts for age group and cancer site (j and k, respectively) in O_{jik} will be suppressed so that other sub- and superscripts can be added without confusion. $O_{i_1 i_2}$ will represent an observed case counted for census tract i_2 of an individual who spent time in i_1 not necessarily equal to i_2. More specifically, $O_{i_1 i_2}(t_1, t_2)$ represents an observed cancer case located in census tract i_2 at the time of the study t_2 of an individual from census tract i_1 at time t_1. This is an unsatisfactory symbolization since it allows only one move per case, an unrealistic and oversimplified view.

The most general representation of O_i would include the possibility for multiple transitions among census tracts and the outside SMSA region. The only restriction would be that the observed case end up in a census tract and not outside of the SMSA at the time of the study. This restriction is obviously required since otherwise the case would not be observed in the study area. This general representation of an observed case would include subscripts for each census tract and/or the area outside of the SMSA in which the case resided: $O_{i_1 i_2 \ldots i_m}$ where i_m is the census tract in which the case is observed (m=1,2,...) and i_1 through i_{m-1} are the other census tracts and/or outside SMSA area in which the case resided prior to being observed in i_m. O_{i_1} would represent a person who was born and resided in a

particular tract until diagnosis. Interest not only lies in the different places of residence but also the lengths of stay in each place. The observed cases would take on new subscripts for lengths of stay, l_1 through l_m: $O_{i_{11_1},i_{11_2},\ldots,i_{ml_m}}$. This is a rather cumbersome symbolization of the observed cancer cases but it takes into account all of the possible moves and the time spent in each place of the individual. It is a starting point from which simplications by assumptions can be made to streamline the analysis.

It is apparent that the identification of cases of cancer that are attributable to asbestos is a difficult process that depends on knowledge about specific aspects of the disease in question and the relationship of agent to disease. Obviously, an $O_{i_{11_i}}$ can be identified as equivalent to an $O_{i_{11_1},i_{21_2}}$ if l_1+l_2 for the second case equals l_1 for the first case and census tracts i_1 and i_2 for the second case have equivalent asbestos levels in their water supplies which are also equal to the asbestos level in the water supply of the census tract i_1 of the first case. In this situation both cases were exposed to the same level of asbestos for the same length of time. This equivalence between cases such as $O_{i_{11_1}}$ and cases that had resided in more than two tracts would also be true as long as the asbestos levels were equivalent throughout and total time spent under these equivalent asbestos levels for both cases were identical.

In the analysis of the data from the asbestos-cancer study an attempt was made to control for the possible biasing effect of migration by removing census tracts with low proportions of population living in the same housing in 1970 as in 1965. This analysis was performed in conjunction with statistical procedures done on data on all available census tracts. In the analysis on the entire data set, the implicit assumption operated that each observed case was $O_{i_{11_1}}$ with l_1 equal to the age of case at diagnosis meaning that each observed case was born in and resided in a particular census tract at least up to the time of diagnosis. It is reasonable to assume that there were a number of cases that fit this situation. These cases and cases of individuals who lived outside of the SMSA for the first x years of their lives and the other years in a census tract are easily aggregated.

SECTION 4

NONSTOCHASTIC MIGRATION MEASURES

Migration is a form of geographic or spatial mobility involving a change of usual residence between clearly defined geographic units. There is international migration which represents movement between countries and there is internal migration which represents movement within a country. In this case we are dealing with internal migration although the models to be reviewed and developed should apply equally well in both cases. The internal migration here might also be considered short distance mobility, because although the movement can involve crossing the country, our main focus is movement within the Standard Metropolitan Statistical Area (SMSA) and out of the SMSA.

There are two major ways of obtaining knowledge about migrants or rate of migration: direct and indirect methods. Direct methods involve obtaining information from the census or surveys in which questions about mobility or prior residence are asked. Indirect methods involve estimating net migration (the difference between in- and out-migrants) from population figures at two censuses or from natural increase (births minus deaths) or intercensal survival rates derived from life tables or comparison of age distributions of successive censuses. These indirect methods are called estimation by the "residual method." The difference between total change in population and change due to natural increase is imputed to net migration. This difference also usually includes net immigration from abroad, however.

The mobility rate of a geographic unit is defined as:

$$\frac{M}{P} k = m = \text{mobility rate}$$

where M = number of movers, P = population at risk of moving, and k is a constant such as 100,000.

Also defined as in-migrant, out-migrant, and net-migrant mobility rates:

$$\frac{I}{P} k = m_i = \text{in-migrant mobility rate}$$

$$\frac{O}{P} k = m_o = \text{out-migrant mobility rate}$$

$$\frac{I-O}{P} k = m_n = \text{net-migrant mobility rate}$$

These rates are calculated over a specified period of time, say, one year.

Although we have defined P as the population at risk of moving, we have not explained how this number is obtained. The population at the beginning of the time period is a possibility. This would not, however, take into account persons born during the migration period who were at risk of moving but not included in the initial population. Persons born or dying during the time period were at risk for only part of that period. Furthermore, if the time period begins at the intercensal date then the initial population can only be estimated approximately. In spite of these problems the population at the beginning of the migration period is clearly preferable to the population at the end of the period.

A possible compromise would be the average of the initial and terminal populations. This would in some sense approximate the person-years at risk in the time period. Another measure that has been suggested is one in which the natural increase and the net migration are eliminated from the population base. Since we might assume that the births, deaths, in- and out-migrations occurred uniformly across the time period, we would add in one-half of the births and in-migrants and subtract out one-half of the deaths and out-migrants to yield:

$$P' = P_0 + \frac{B-D + I-O}{2}$$

where P' is the adjusted population at risk, P_0 is the initial population, B is the total births, D is the total deaths, I is the total in-migrants, and O is the total out-migrants. This type of measure would depend on the existence of a population register in order to obtain a count of the out-migrants.

For the various mobility rates we need to decide the area from which the population base is to be chosen. In the case of the out-migration rate the population at risk is the area itself. For in-migration the population at risk is that of the balance of the country (under internal migration). Net migration rates must be the difference of in and out migration rates that have the same population base.

The indirect measures of migration presented here are (1) the national growth rate method, and (2) the residual method, comprising (a) the vital statistics method and (b) the survival rate method.

The national growth rate method is crude but very commonly used. Under the national growth rate method the estimated migration rate is:

$$m_i = \left[\frac{P_i'-P_i^0}{P_i^0} - \frac{P_T'-P_T^0}{P_T^0} \right] k \quad , \qquad (1)$$

where m_i is the estimated migration rate, P_T^0 and P_T' are national population at beginning and end of the intercensal period, $P_1^0, P_2^0, \ldots, P_i^0$ are the populations

of the geographic subdivisions at the beginning of the period and $P'_1, P'_2, \ldots,$ their populations at the end of the period. This rate is customarily multiplied by a constant such as 100 or 1,000.

Thus, for a geographic unit, a rate of growth greater than the national average is interpreted as net in-migration, and a rate less than the national average as net out-migration. The same procedure can be applied to specific age-sex groups. This method yields an estimate of the rate of internal migration for each geographic subdivision on the assumption that rates of natural increase and of net immigration from abroad are the same for all parts of the country. It requires no vital statistics.

A slightly different version of this method compares the rate of growth in the geographic subdivision with the rate of natural growth of the nation. This method assumes that the rate of natural increase is the same throughout the country.

Here,

$$m_i = \left[\frac{P'_i - P^0_i}{P^0_i} - \frac{B_T - D_T}{P^0_T} \right] k \ , \qquad (2)$$

where the last term is the national rate of natural increase. Where net immigration is negligible, formulas (1) and (2) yield virtually the same results. When either net immigration or natural increase is not available, it can be estimated by subtraction of its counterpart from total increase. One danger of estimating either natural increase or net immigration by subtraction of its counterpart from total increase is that there may be considerable under-registration or that one census may be more nearly complete than the other.

There are several varieties of the residual method. The vital statistics method is conceptually simpler than the survival rate method, but it is not feasible where records are poor. The vital statistics method requires virtually complete registration of both births and deaths during the intercensal period. If, however, the extent of under-registration has been estimated, the correct vital statistics should be used. The births and deaths should also be by date of occurrence (not date of registration) and ordinarily by place of residence (not place of occurrence).

The vital statistics method is almost always used to estimate net migration for the total population of an area, i.e., for all age-sex groups combined. The simplest equation is a form of the "balancing equation":

$$M = (P' - P^0) - (B - D) \ , \qquad (3)$$

or net migration equals population change during the intercensal period minus natural increase during that period. The "intercensal component method" is

a means of estimating the residual, i.e., net migration. The result is the difference between the total number of persons moving into an area during a given intercensal period and the number moving out. This estimate reflects both the in-migrants and out-migrants who died before the second census. The net migration obtained for a given geographic subdivision is that with respect to all other areas and thus represents net international migration combined with net internal migration. This method may be used to estimate net migration for a sex-race, or nativity group or a group defined by any other characteristic that is invariant over time, provided that the population and vital statistics are published by the characteristic. The main issue concerning the vital statistics method is not the theoretical validity of formula (3) but rather the accuracy of the data for the terms on the right hand side of this equation. In other words, the main concern is with the amount of error in estimating M from P', P^0, B, and D. Of further interest is a comparison of the vital statistics method with other methods such as the survival rate method to ascertain accuracy of estimation when applied to actual statistics.

Estimates of net migration by the vital statistics residual method are likely to appear after a census rather than annually since annual population estimates made by the component method require a prior estimate of net migration. Estimates for geographic subdivisions, then, seldom appear in national statistical yearbooks but are more likely to appear in special reports such as "Current Population Reports" published by the US Bureau of the Census. These estimates might also be published by state and/or city governments. Furthermore, such estimates are sometimes made by demographers at universities, special commissions, or international research centers.

An important source of error in these estimates of net migration is a change in the boundaries of geographic area or areas in question. Unless the population figures and data on natural increase can be adjusted to represent a constant area (ordinarily the present rather than the original area), the estimates will reflect the population in the transferred territory (the added or deleted area across time) as well as net migration. The constant area problem represents itself in the case of census tract comparability between censuses. In the case of changing boundaries, tracts should be combined at one or both census dates to obtain an identical area before the change in population is computed.

As mentioned above, variants of the residual method include in their estimates of net migration net immigration from abroad -- unless a specific attempt is made to remove that component. One approach to solve this problem is to confine the computations to the native population. In the case of the vital statistics method, this approach requires that the native population be available separately from the two censuses and that deaths also be tabulated by nativity. The inputs and methods of this approach are as follows:

$$P_2 - P_1 - B + D = \text{Estimated Net Migration}$$

where P_2 = native population at second census, P_1 = native population at first census, B = total births in intercensal period, and D = deaths of

natives in intercensal period. Native here means born in the country, not just born in the geographic subdivision in question. Births are all native by definition, and babies born during the intercensal period are included in the native population counted at the end of the period.

Another approach is to try to allow for net immigration from abroad by subtraction. In this case, one must know not only the total immigration but also the area (county, census tract, etc.) of intended residence of the immigrants and the area of last residence of the emigrants. It is necessary to assume that immigrants went to their announced area of destination and stayed until the next census and that emigrants were counted at the preceding census in the area given as that of last residence.

Since the estimates of net migration obtained through these methods depend on a residual obtained by subtraction, relatively moderate errors in population counts or in statistics of births or deaths produce much larger percentage errors in the migration estimates. However, because the population counts at two successive censuses may represent about the same amount of under-enumeration these errors would be offsetting. If the under-enumeration varies from census to census, it would then be necessary to know the under-enumeration of the second census relative to that of the first rather than the absolute under-enumeration at each census.

The vital statistics method is adaptable to making estimates of net migration by age. For an age group alive at the earlier census, net migration may be expressed as follows:

$$M_a = P^t_{a+t} - P^0_a + D ,$$

where M_a is the net migration of the cohort aged a at the earlier census, P^t_{a+t} is the population aged a+t enumerated at the later census (t years later later), P^0_a is the population aged a enumerated at the earlier census, and D is the deaths to the cohort over the intercensal period. For example, net migration for persons 10 years old at the later census is obtained from the formula $M_{0 \text{ to } 9} = P^t_{0 \text{ to } 9} - B + D_B$, where B is the number of resident births during the decade and D_B is the number of resident deaths to persons born during the decade.

The survival rate residual method is a commonly used variant which yields estimates of net migration by age and sex without nearly as much labor as in the vital statistics residual method. The basic formula for estimating net migration is:

$$M_{x+t} = P^t_{x+t} - sP^0_x \qquad (4)$$

where x is an age or age group, t is the interval in years between censuses, P^0_x is the population aged x at the first census, P^t_{x+t} is the population at

the next census at x+t, and s is the survival rate. The second term on the right-hand side of this formula is the expected population at the second census had there been no net migration.

Survival rates are obtained from either life tables or censuses. The former are derived from a life table, if possible, for the same geographic area and time period to which the estimate to net migration applies. If the estimates were to be made at the county level and life tables were unavailable for counties then one could use a recent life table for its state or region. For a city, for example, the life table for the national urban population might be used. The census survival rate represents the ratio of the numbers in the same national cohort at successive censuses. The objective is to approximate a closed population. This method requires no life table, or no vital statistics, and has the advantage of avoiding the effects of some of the errors in the population estimates.

Unlike the vital statistics method, neither variant of the survival rate method measures net migration exactly even when there are no errors in the underlying population and vital statistics. Formula (4) estimates the number of net migrants who were living at the end of the period. This difference arises because the implicit number of deaths is not necessarily equivalent to the numbers of deaths to residents of the given area during the given period, that is, deaths to non-migrants plus deaths to in-migrants.

There are several ways of applying the survival rates to this method. Formula (4) is known as the forward survival rate method. The reverse survival rate method divides the number in the age group at the end of the intercensal period by the survival rate yielding:

$$M'_{x+t} = \frac{P^{t}_{x+t}}{s} - P^{0}_{x} \quad . \tag{5}$$

The two methods always yield different results but are related by $M'_{x+t} = M_{x+t}/s$. The results of the forward and reverse methods may be averaged to yield:

$$M''_{x+t} = \frac{M_{x+t} + M'_{x+t}}{2} \quad .$$

The average implies a more meaningful assumption regarding the timing of net migration than does either the forward or reverse method. The amount of difference between the migration estimates from the forward and reverse methods depends on the amount of net migration and on the level of the survival rate. The percentage difference is a function of only the survival rates.

Under the life table survival rate method, the survival rate for a 5 year age group and 10 year period is

$$s_x = {}_5L_{x+10} / {}_5L_x \qquad (6)$$

where s_x is the survival rate in the age group, ${}_5L_{x+10}$ is the number of person years that would be lived within the 5-year age group (x+10 to x+15) by the cohort of the number of births assumed, and ${}_5L_x$ is the number of person-years that would be lived within the 5-year age group (x to x+5) by the cohort of the number of births assumed. For example, the estimated net migration is:

$$M_{20-24} = P^{10}_{20-24} - \frac{{}_5L_{20}}{{}_5L_{10}} \times P^{0}_{10-14} \quad .$$

If there is an open-ended interval, say 75 years old and over, in the census distribution, then $s_{70} = T_{75}/T_{70}$ where s_{70} is the survival rate for this age group, T_{75} is the total number of person-years that would be lived after the beginning of the 75 years old and over age group by the cohort of the number of assumed births, and T_{70} is the total number of person-years that would be lived after the beginning of the 70-74 year age group by the cohort of the number of assumed births.

Ideally, a life table covering the full intercensal period would be used to compute these survival rates. If, however, there were tables available only for the two census dates, the survival rates could be computed from both life tables and the results averaged to yield a more representative view of survival over the decade than either of the two tables taken individually.

With an abridged table and when the intercensal period is not 5 or 10 years, interpolation methods would be needed. Survivors will appear in conventional 5-year age groups, and therefore, should be redistributed into conventional 5-year age groups so that they may be compared with the age groups of the second census. Estimates of net migration obtained by subtraction would then be in terms of conventional age groups.

The census survival rate method is employed by taking the ratio of the population aged x+n at the second census to that aged x at the first census, where the censuses are taken n years apart. Thus,

$$s^n_x = \frac{P^{t+n}_{x+n}}{P^{t}_{x}} \qquad (7)$$

where t is the date of the first census. We want this rate to reflect mortality but not migration. Therefore, census survival rates have to be based on national population statistics. If there is appreciable external migration then it is preferable to base this rate on the native population as counted in the two national censuses. It is then permissible to apply this rate to the total population figures for local areas so as to include the net

migration of the former immigrants in the estimates. This assumes that foreign born and native populations have the same survival rates. There are also two other important assumptions made in this method: (1) true survival rates are the same for geographic subdivisions as for the nation and (2) the pattern of relative errors in the census age data is also the same. The first of these two assumptions also operates when we use a life table for a larger area containing the area in question or when we use a model life table (estimated life table based on a wider range of data due to poor reporting in the area in question).

Estimates of net migration by age and sex made by the survival rate method are sometimes adjusted to add to an estimate for the total population made by the vital statistics method. The latter gives a theoretically exact measure of net migration whereas the former does not. The accuracy of estimates made by these two methods depends upon a number of empirical considerations.

In summary, the residual method cannot be used to estimate gross immigration or out-migration or migration streams. The migration period must be the intercensal period. In any of the variants of this method, the net migration can be estimated for a fixed area, for a group defined by an unchanging characteristic (sex, race, nativity, etc.) or for a group defined by a characteristic that changes in a fixed way with time (age).

Place-of-birth statistics, obtained from either censuses or sample surveys, can be used to measure migration. Unlike estimates of migration from the residual method which are limited to net movements, place-of-birth data can represent in-migrants, out-migrants, and specific streams. These data do not allow for estimation of intra-area mobility (short distance movement) nor do they take account of intermediate migrations (moves between the time of birth and the time of the census); and persons who have returned to live in their area of birth appear as nonmigrants. In other words, these statistics do not indicate the total number of persons who have moved from the area in which they were born to other areas, or to any specific area, during any given period of time.

If census statistics are not tabulated by age, the time that a move occurred is unknown. With an age tabulation, the only specification is that given by age itself. That is, a migrant 25 years old must have moved within the 25 years preceding the census. Thus, the older the migrant the less we know about the date of the move and the more likely that the migrant made intervening moves between other areas of the same class. Statistics tabulated by age for two successive censuses greatly enhance the data but are not fully adequate for measuring migration in an intercensal period.

Statistics on birth place (state, province, etc.) relate to the native population of the country. Therefore, the internal migration of the foreign-born population subsequent to its immigration is not included.

In computing migration rates from place-of-birth statistics, there are appropriate situations for using either the population at origin or the population at destination. In either case, the population at the time of

the census must be used. The interregional migration rate is defined by:

$$m_r = \frac{\sum_i \sum_j N_{ij} - \sum_{i=j} \sum N_{ij}}{N} \times 100$$

where N is the native population; subscript i refers to the region of enumeration and subscript j to the region of birth; and N_{ij} is the number of natives living in region i and born in region j, including those living in the region of birth (i=j). Thus, $\sum_i \sum_j N_{ij} = N$. $\sum_{i=j} \sum N_{ij}$ is the number of natives living in the region of birth. $N_{i \neq j}$ is an interregional migration stream, which we may call M_{ij}. This leads to an alternative expression of the interregional migration rate:

$$m_r = \frac{\sum_i \sum_j M_{ij}}{N} \times 100 \quad .$$

The in-migration rate for a region is:

$$m_1^I = \frac{\sum_j M_{1j}}{N_1} \times 100$$

where M_{1j} is the migrants living in region 1 who were born in region j; and N_1 is the native population enumerated in region 1. Note that $\sum_j M_{1j} = \sum_j N_{1j} - N_{11} = N_1 - N_{11}$.

The out-migration rate for a region is:

$$m_1^O = \frac{\sum_i M_{i1}}{N_1'} \times 100$$

where $\sum_i M_{i1}$ is the migrants from region 1 to the i-th region and N_1' is the total population born in region 1. Again, $\sum_i M_{i1} = \sum_i N_{i1} - N_{11} = N_1' - N_{11}$.

The net migration rate for a region is:

$$m_1^n = \frac{\sum_j M_{1j} - \sum_i M_{i1}}{N_1} \times 100 \quad .$$

If the out-migration rate is computed using the native population living in the region as the base, the net migration rate could be obtained by subtracting the out-migration rate from the in-migration rate.

The turnover rate which is defined as the sum of the in- and out-migration rates is:

$$m_1^T = \frac{\sum_j M_{1j} + \sum_i M_{i1}}{N_1} \times 100 = m_1^I + m_1^O \quad .$$

The birth-residence index is the net gain or loss for an area through interarea migration. In other words, it is the net effect of lifetime migration upon the surviving population. The formula for area 1 is:

$$BR_1 = \sum_j M_{1j} - \sum_i M_{i1} = I_1 - O_i$$

which is the numerator of the net migration rate. Thus, the sum of birth-residence indices across all areas of a country must be equal to zero, or

$$\sum_i BR_i = 0 \quad .$$

As with any measure of net migration, the results may reflect qualitatively different patterns of in- and out-migration over time. A net migration of zero may represent a lack of in- and out-migration, or net in-migration in recent decades offset by net out-migration in earlier decades, or net out-migration in recent decades balanced by heavier net in-migration in earlier decades, or more complex patterns. A given amount of migration in a recent decade has a similar effect to that of a larger migration at an earlier decade because of different proportions of survivors from migrants of the two decades; and the analysis is further complicated by differences in the age distributions of the migrants in the two decades and by changing mortality conditions.

The intercensal change in the birth-residence index is a measure of the difference in the birth-residence indices of two consecutive census periods for a given state (or group of states) and is intended to estimate the net migration during the latter decade. This measure is defined as:

$$BR_{i2} - BR_{i1} = (I_{i2} - I_{i1}) - (O_{i2} - O_{i1})$$

where the subscripts 2 and 1 indicate the first and second censuses, respectively, and the subscript i refers to the area in question.

The gross intercensal interchange of population is the sum of the absolute values of the change during an intercensal period in the number of nonresident natives of an area and the change in the number of resident natives of other states. This is a measure of gross interstate migration, or of population turnover, for a given state. This measure is defined as:

$$GIIP = I_{i2} - I_{i1} + O_{i2} - O_{i1} \quad .$$

The intercensal change in the birth-residence index ($BR_{i2}-BR_{i1}$) can be refined to allow for intercensal mortality as follows:

$$BR_{i2}-BR_{i1} = (I_{i2}-S^{I}I_{i1}) = (O_{i2}-S^{O}I_{i1})$$

where S^{I} and S^{O} are the intercensal survival rates for the life-time in-migrants and out-migrants as counted at the earlier census, respectively. The two terms of this modified formula give net migration among persons born outside the area and inside the area, respectively.

Place-of-birth data are used to compute the intercensal survival rates. When place-of-birth data are tabulated by age for neither of two successive censuses, it is possible to use only an overall intercensal or life-table survival rate (T_{10}/T_{0}). It is presumed that migrants had the age distributions and age-specific mortality rates of the general (life table) population. When age is cross-tabulated at only one census (usually the later of two censuses), the intercensal survival rate for a period of k years is computed by getting the ratio of the population k years old and over at the second census to that of all ages at the first census. An underlying assumption here is that persons born in a given area will have the same survival rates regardless of their subsequent migration histories. When place-of-birth data are tabulated by age at both censuses, the strategy is similar to that used when place-of-birth data are tabulated by age for only the later census. The computations are carried out for each age cohort separately.

The applicability of these procedures to the analysis of migration related to census tracts is dependent upon the availability of data to secure and compute the components of these procedures. The national growth rate method applies to census tract migration analysis since it requires only population counts by census tracts and for the entire country. These figures are readily available from the national census. The vital statistics residual method cannot be directly computed for census tracts since it requires population counts which are available and survival rates (based on life-table or census data) which can be estimated. Life tables for census tracts may be unavailable, therefore, eliminating the possiblity of using the life-table survival rate method but the census survival rate method can be employed since this rate is based on available population counts. Place-of-birth methods do not apply since statistics are not recorded for census tract of birth.

SECTION 5

STOCHASTIC MODELS OF MIGRATION

In the first model that we shall consider, the Poisson process, the probability that an event occurs during the time interval $(t,t+\Delta)$ is: $\eta\Delta+o(\Delta)$, where the constant η does not depend on t or on the number of events occuring in $(0,t)$; and the probability of no change in $(t,t+\Delta)$ is: $1-\eta\Delta-o(\Delta)$. In this case, an event refers to the act of migration by an individual into the area of interest and $o(\Delta)$ represents any function of Δ which tends to zero faster than Δ, that is, any function such that: $o(\Delta)/\Delta \rightarrow 0$ as $\Delta \rightarrow 0$.

Therefore, we can write the probability of getting k immigrants in the area of interest during the time interval $(0,t+\Delta)$ as:

$$p_k(t+\Delta) = p_k(t)(1-\eta\Delta) + p_{k-1}(t)\eta\Delta + o(\Delta) \tag{1}$$

and, for k=0,

$$p_0(t+\Delta) = p_0(t)(1-\eta\Delta) + o(\Delta) \ . \tag{2}$$

Transposing $p_k(t)$ to the left side of (1), dividing the resulting equation through by Δ, and passing to the limit as $\Delta \rightarrow 0$, we find that the probabilities satisfy the system of differential equations:

$$\frac{d}{dt} p_k(t) = -\eta p_k(t) + \eta p_{k-1}(t) \ , \qquad \text{for } k \geq 1 \ , \tag{3}$$

and, for k=0,

$$\frac{d}{dt} p_0(t) = -\eta p_0(t) \ , \tag{4}$$

with initial conditions,

$$p_0(0) = 1, \quad p_k(0) = 0, \quad k \geq 1 \ . \tag{5}$$

Integrating (4) and using initial condition $p_0(0) = 1$ yields:

$$p_0(t) = e^{-\eta t} \ . \tag{6}$$

Setting k=1 in (3) and multiplying through by $e^{\eta t}$ gives:

$$e^{\eta t} \frac{d}{dt} p_1(t) + \eta e^{\eta t} p_1(t) = \frac{d}{dt} e^{\eta t} p_1(t) = \eta \quad . \tag{7}$$

Integrating (7) and using initial condition $p_1(0) = 0$, we have:

$$p_1(t) = \eta t e^{-\eta t} \quad . \tag{8}$$

Repeated applications of the same procedure yield the general formula:

$$p_k(t) = \frac{e^{-\eta t}(\eta t)^k}{k!} \quad , \quad k=0,1,\ldots \quad . \tag{9}$$

This is the Poisson distribution with parameter ηt.

The application of this process to migration is that $\eta\Delta$ gives the probability of an in-migrant in a time interval of length Δ. In the case of the asbestos-cancer study, we are interested in migration into census tracts. Therefore, we would have a system of Poisson processes:

$$p_{ki}(t) = \frac{e^{-\eta_i t}(\eta_i t)^k}{k!} \quad , \quad k=0,1,\ldots \quad . \tag{10}$$

The subscript i refers to the specific census tract in question (i=1,..., 722).

It may be more realistic to assume that η is a function of time, $\eta(t)$, so that the probability of migrating varies with t. Thus, the probability of the occurrence of k in-migrants in (0,t) for the i-th census tract is:

$$p_{ki}(t) = \frac{\exp\{-\int_0^t \eta_i(\tau)d\tau\}\left[\int_0^t \eta_i(\tau)d\tau\right]^k}{k!} \quad , \quad k=0,1,\ldots \quad . \tag{11}$$

The expected value of $X_i(t)$, the number of in-migrants in (0,t) for census tract i, would be

$$E[X_i(t)] = \int_0^t \eta_i(\tau)d\tau \quad . \tag{12}$$

Up to this point we have considered only in-migration in our analysis. Now we include the possibility for both immigration and emigration. We do this by constructing a process with parameters representing the instantaneous probabilities of in- and out-migration. We let these parameters depend on k, the number of individuals in the area in question. Note the difference

between the interpretation of k in this case versus its meaning under the Poisson process. Here, k refers to the number of individuals in an area; whereas under the Poisson process, k refers to the number of in-migrants in an area.

Let $\eta_k(t)\Delta + o(\Delta)$ = probability of an in-migrant in $(t,t+\Delta)$, $\mu_k(t)\Delta + o(\Delta)$ = probability of an out-migrant in $(t,t+\Delta)$, and $1-\eta_k(t)\Delta - \mu_k(t)\Delta - o(\Delta)$ = probability of no change in $(t,t+\Delta)$. Therefore,

$$p_k(t+\Delta) = p_k(t)[1-\eta_k(t)\Delta - \mu_k(t)\Delta] + p_{k-1}(t)\eta_{k-1}(t)\Delta + p_{k+1}(t)\mu_{k+1}(t)\Delta + o(\Delta) \quad , \tag{13}$$

yielding the system of differential equations

$$\frac{d}{dt} p_k(t) = -[\eta_k(t)+\mu_k(t)]p_k(t) + \eta_{k-1}(t)p_{k-1}(t) + \mu_{k+1}(t)p_{k+1}(t) \quad , \quad k \geq 1 \tag{14}$$

$$\frac{d}{dt} p_0(t) = -[\eta_0(t)+\mu_0(t)]p_0(t) + \mu_1(t)p_1(t) \quad , \tag{15}$$

for k=0, with initial conditions,

$$p_{k_0}(0) = 1, \quad p_k(0) = 0, \quad k \neq k_0 \quad , \tag{16}$$

where k_0 is the population size at time 0.

Under the above instantaneous probabilities, the assumption operates that migration is dependent on both time and the number of persons in an area at a particular time. This would be realistic for out-migration since the probability of migrating from a place obviously depends on the number of persons in that place. However, for in-migration, this is not necessarily the case. The probability of in-migration would depend on the population of people outside the area in question but not on the number of people inside that area. Therefore, a general form of the in- and out-migration process with realistic assumptions concerning the instantaneous probabilities would be based on:

$\eta(t)\Delta + o(\Delta)$ = probability of an in-migrant in $(t,t+\Delta)$

and

$\mu_k(t)\Delta + o(\Delta)$ = probability of an out-migrant in $(t,t+\Delta)$.

This yields

$$p_k(t+\Delta) = p_k(t)[1-\eta(t)\Delta - \mu_k(t)\Delta] + p_{k-1}(t)\eta(t)\Delta + p_{k+1}(t)\mu_{k+1}(t)\Delta + o(\Delta) \tag{17}$$

and

$$\frac{d}{dt} p_k(t) = -[\eta(t)+\mu_k(t)]p_k(t) + \eta(t)p_{k-1}(t) + \mu_{k+1}(t)p_{k+1}(t) , \quad k \geq 1 , \tag{18}$$

$$\frac{d}{dt} p_0(t) = -[\eta(t)+\mu_0(t)]p_0(t) + \mu_1(t)p_1(t) , \quad k=0 , \tag{19}$$

with initial conditions,

$$p_{k_0}(0) = 1, \quad p_k(0) = 0, \quad k \neq k_0 . \tag{20}$$

It will be assumed that the reduction in population due to out-migration is linear in the number of people in the area in question. That is,

$$\mu_k(t) \to k\mu(t) .$$

We will solve this process by noting that the population at time t is the sum of two random variables:

$$X(t) = Y(t) + Z(t) , \tag{21}$$

where X(t) is the population at time t, Y(t) is the number of survivors of the initial k_0 people, and Z(t) is the number of people who enter the population during the interval (0,t) and are still present at time t. Since the survivors of the initial population are independent of the survivors of the in-migrants, Y(t) and Z(t) are independent random variables and therefore, the probability generating function of X(t) is the product of the probability generating functions of Y(t) and Z(t):

$$g_X(s;t) = g_Y(s;t) \times g_Z(s;t) \ . \tag{22}$$

The probability generating function for $Y(t)$, $g_Y(s;t)$ can be solved by the method of probability generating functions. The probability that an individual out-migrates during the time interval $(t,t+\Delta)$ is: $\mu(t)\Delta + o(\Delta)$; and the probability of remaining in $(t,t+\Delta)$ is: $1-\mu(t)\Delta - o(\Delta)$. Therefore, the probability of having k individuals in an area at time $t+\Delta$ is:

$$p_k(t+\Delta) = p_k(t)[1-k\mu(t)\Delta] + p_{k+1}(t)(k+1)\mu(t)\Delta + o(\Delta) \tag{23}$$

and, for k=0,

$$p_0(t+\Delta) = p_0(t) + p_1(t)\mu(t)\Delta + o(\Delta) \ . \tag{24}$$

Transposing $p_k(t)$ to the left side of (23), dividing the resulting equation through by Δ, and passing to the limit as $\Delta \to 0$. we find that the probabilities satisfy the system of differential equations:

$$\frac{d}{dt} p_k(t) = -k\mu(t)p_k(t) + (k+1)\mu(t)p_{k+1}(t) \ , \quad \text{for } k \geq 1 \ , \tag{25}$$

and, for k=0,

$$\frac{d}{dt} p_0(t) = \mu(t)p_1(t) \ , \tag{26}$$

with initial conditions,

$$p_{k_0}(0) = 1, \quad p_k(0) = 0, \quad k \neq k_0 \ . \tag{27}$$

The probability generating function of the probability distribution $p_k(t)$ is defined as:

$$g_Y(s;t) = \sum_{k=0}^{\infty} p_k(t)s^k \ , \quad |s| \leq 1, \ 0 \leq t \leq \infty \ , \tag{28}$$

and

$$g_y(s;0) = s^{k_0} \ . \tag{29}$$

Differentiation of the probability generating function under the summation sign yields:

$$\frac{\partial}{\partial t} g_Y(s;t) = \sum_{k=0}^{\infty} [\frac{d}{dt} p_k(t)]s^k \quad . \tag{30}$$

We take the derivative of $g_Y(s;t)$ with respect to t and use the system (25) to obtain the differential equation:

$$\frac{\partial}{\partial t} g_Y(s;t) = -s\mu(t) \frac{\partial}{\partial s} g_Y(s;t) + \mu(t) \frac{\partial}{\partial s} g_Y(s;t)$$

$$= \mu(t)(1-s) \frac{\partial}{\partial s} g_Y(s;t) \quad . \tag{31}$$

The auxiliary equations are:

$$\frac{dt}{-1} = \frac{ds}{\mu(t)(1-s)} , \quad dg_Y(s;t) = 0 \quad . \tag{32}$$

The first equation may be written as:

$$-\mu(t)dt = \frac{ds}{1-s} \iff \mu(t)dt = d\ \ln(1-s) \iff (1-s)\ e^{-\int_0^t \mu(\tau)d\tau} = \text{constant.} \tag{33}$$

The second auxiliary equation implies that:

$$g_Y(s;t) = \text{constant} .$$

Therefore, the general solution of (31) is

$$g_Y(s;t) = \Phi(1-s)\ e^{-\int_0^t \mu(\tau)d\tau} \tag{34}$$

where Φ is an arbitrary differentiable function. To obtain the particular solution of $g_Y(s;t)$ corresponding to the initial condition

$$g_Y(s;0) = s^{k_0} ,$$

we set t=0 in (34) to obtain:

$$s^{k_0} = \Phi(1-s) = g_Y(s;0) \quad . \tag{35}$$

Equation (35) holds true at least for all s with $|s| < 1$; hence, for any θ such that $|1-\theta| < 1$,

$$\Phi(\theta) = (1-\theta)^{k_0} \quad . \tag{36}$$

Letting

$$\theta = e^{-\int_0^t \mu(\tau)\,d\tau} (1-s) \tag{37}$$

we get

$$g_Y(s;t) = \left[1-(1-s)e^{-\int_0^t \mu(\tau)\,d\tau}\right]^{k_0} \quad . \tag{38}$$

Therefore, the number of survivors of the initial k_0 people, $Y(t)$, is a binomial random variable with parameters k_0 and

$$e^{-\int_0^t \mu(\tau)\,d\tau} \quad .$$

To find the probability generating function of $Z(t)$, the number of in-migrants who survived to time t, we consider an individual that immigrates into the population in $(\tau,\tau+\Delta)$ with probability $\eta(\tau)\,d\tau$ and survives to t with probability

$$e^{-\int_\tau^t \mu(\xi)\,d\xi} \quad , \quad 0 \leq \tau \leq t \quad .$$

Therefore, the probability that an individual immigrates into the population in $(\tau,\tau+\Delta)$ and survives to t is:

$$[\eta(\tau)\,d\tau]\, e^{-\int_\tau^t \mu(\xi)\,d\xi} \quad . \tag{39}$$

Summing (39) over all possible values of τ, we obtain

$$\int_0^t \eta(\tau)\, e^{-\int_\tau^t \mu(\xi)\,d\xi}\, d\tau \quad . \tag{40}$$

Since Z(t) is the total number of events described as surviving immigrants in (0,t), Z(t) has a Poisson distribution with expectation (40). Therefore

$$g_Z(s;t) = \exp\{-(1-s)\int_0^t \eta(\tau)\, e^{-\int_\tau^t \mu(\xi)\,d\xi}\, d\tau\} \quad . \tag{41}$$

Therefore, since the probability generating function of X(t), the population at time t, is the product of the probability generating functions of Y(t) and Z(t); we get:

$$g_X(s;t) = \left\{1-(1-s)\, e^{-\int_0^t \mu(\tau)\,d\tau}\right\}^{k_0} \exp\left\{-(1-s)\int_0^t \eta(\tau)\, e^{-\int_\tau^t \mu(\xi)\,d\xi}\, d\tau\right\} \tag{42}$$

Another consideration in the analysis of population growth is births and deaths. One of the most basic stochastic models that includes parameters for births, deaths, and migration is the Kendall process. In this process the instantaneous probabilities take the form:

$\lambda\Delta + o(\Delta)$ = probability an individual will reproduce one individual in $(\tau,\tau+\Delta)$,

$\mu\Delta + o(\Delta)$ = probability of death of an individual in $(\tau,\tau+\Delta)$,

$\eta\Delta + o(\Delta)$ = probability of an individual entering the population by migration in $(\tau,\tau+\Delta)$.

This process is such that λ and μ apply independently to each and every member of the population and η acts independently of the population. Each member of the population produces a birth with probability λ and dies with probability μ; and a migrant enters the population with probability η regardless of the size of the population. Furthermore, we can allow μ to represent death and/or emigration yielding a process that includes births, deaths, in-migrants, and out-migrants. We need separate parameters for birth and immigration since birth is dependent on the number of persons in the study area (more specifically, the number of females), while immigration is not necessarily dependent on the number of people in the study area.

We write the probability of k individuals in the study area at time $(t+\Delta)$ as

$$p_k(t+\Delta) = p_k(t)[1-(k\lambda+\eta+k\mu)\Delta - o(\Delta)] + p_{k-1}(t)[(\eta+(k-1)\lambda)\Delta + o(\Delta)]$$
$$+ p_{k+1}(t)[(k+1)\mu\Delta + o(\Delta)] \quad , \quad \text{for } k \geq 1 \tag{43}$$

and

$$p_0(t+\Delta) = p_0(t)[1-\eta\Delta-o(\Delta)] + p_1(t)[\mu\Delta+o(\Delta)] , \quad \text{for } k=0 .$$

Rewriting (43), dividing by Δ, and passing to the limit as $\Delta \to 0$, we get the following system of differential equations:

$$\frac{d}{dt} p_k(t) = -p_k(t)[k(\lambda+\mu)+\eta] + p_{k-1}(t)[(k-1)\lambda+\eta] + p_{k+1}(t)[(k+1)\mu] , \quad \text{for } k \geq 1 , \tag{44}$$

$$\frac{d}{dt} p_0(t) = -p_0(t)\eta + p_1(t)\mu , \quad \text{for } k=0 \tag{45}$$

with initial conditions,

$$p_{k_0}(0) = 1, \quad p_k(0) = 0, \quad k \neq k_0 . \tag{46}$$

The solution of this differential equation can also be obtained by the method of probability generating functions. Taking the derivative of the probability generating function with respect to t and using the system (44), we obtain the partial differential equation:

$$\frac{\partial}{\partial t} g_X(s;t) = -(1-s)(s\lambda-\mu) \frac{\partial}{\partial s} g_X(s;t) - \eta(1-s) g_X(s;t) . \tag{47}$$

The auxiliary equations are:

$$\frac{dt}{1} = \frac{ds}{(1-s)(s\lambda-\mu)} = \frac{dg_X(s;t)}{-\eta(1-s) g_X(s;t)} . \tag{48}$$

By the method of partial fractions, the second term of the first equation may be written as:

$$\frac{ds}{(1-s)(s\lambda-\mu)} = \left[\frac{\lambda}{(\lambda-\mu)(s\lambda-\mu)} + \frac{1}{(\lambda-\mu)(1-s)}\right] ds .$$

Therefore, the first equation may be rewritten as:

$$\frac{dt}{1} = \frac{\lambda}{(\lambda-\mu)(s\lambda-\mu)} ds + \frac{1}{(\lambda-\mu)(1-s)} ds \iff (\lambda-\mu)dt = d[\ln(s\lambda-\mu)] - d[\ln(1-s)] .$$

Collecting terms and integrating yields:

$$\frac{(1-s)}{(s\lambda-\mu)} e^{(\lambda-\mu)t} = \text{constant} . \tag{50}$$

The second auxiliary equation is

$$\frac{ds}{(1-s)(s\lambda-\mu)} = \frac{dg_X(s;t)}{-\eta(1-s)g_X(s;t)} \iff \frac{-\eta ds}{(s\lambda-\mu)} = \frac{dg_X(s;t)}{g_X(s;t)} . \tag{51}$$

Integrating yields:

$$(s\lambda-\mu)^{\eta/\lambda} g_X(s;t) = \text{constant} . \tag{52}$$

Therefore, the general solution to the partial differential equation is:

$$g_X(s;t) = (s\lambda-\mu)^{-\eta/\lambda} \phi\{\frac{(1-s)}{(s\lambda-\mu)} e^{(\lambda-\mu)t}\} \tag{53}$$

where ϕ is an arbitrary differentiable function. To obtain the particular solution of $g_X(s;t)$ corresponding to the initial condition ($g_X(s;0) = s^{k_0}$), we set t=0 in (53) to obtain:

$$g_X(s;0) = s^{k_0} = (s\lambda-\mu)^{-\eta/\lambda} \phi\{\frac{(1-s)}{(s\lambda-\mu)}\} . \tag{54}$$

Equation (54) holds true at least for all s with $|s| < 1$; hence, for any θ such that $|(1+\mu\theta)/(1+\lambda\theta)| < 1$,

$$\phi(\theta) = \{\frac{(1+\mu\theta)}{(1+\lambda\theta)}\}^{k_0} (\{\frac{(1+\mu\theta)}{(1+\lambda\theta)}\} \lambda-\mu)^{\eta/\lambda} . \tag{55}$$

Letting

$$\theta = \frac{(1-s)}{(s\lambda-\mu)} e^{(\lambda-\mu)t} \tag{56}$$

yields

$$g_x(s;t) = \left\{\frac{1 + \mu\frac{(1-s)}{(s\lambda-\mu)} e^{(\lambda-\mu)t}}{1 + \lambda\frac{(1-s)}{(s\lambda-\mu)} e^{(\lambda-\mu)t}}\right\}^{k_0} \left\{\left[\frac{1 + \mu\frac{(1-s)}{(s\lambda-\mu)} e^{(\lambda-\mu)t}}{1 + \lambda\frac{(1-s)}{(s\lambda-\mu)} e^{(\lambda-\mu)t}}\right]\lambda-\mu\right\}^{\eta/\lambda} (s\lambda-\mu)^{-\eta/\lambda} . \tag{57}$$

Algebraic manipulation of this expression yields:

$$g_X(s;t) = \left(\frac{p}{1-qs}\right)^{\eta/\lambda}\left[\frac{1 + \left(\frac{q}{p} - \gamma\right)(1-s)}{1 + \frac{q}{p}(1-s)}\right]^{k_0} \tag{58}$$

where

$$\gamma = e^{(\lambda-\mu)t}, \quad p = \begin{cases} \dfrac{\lambda-\mu}{\gamma\lambda-\mu}, & \lambda\neq\mu \\ (1+\lambda t)^{-1}, & \lambda=\mu \end{cases}, \quad \text{and } q = 1-p .$$

The expression for $g_X(s;t)$ in (58) can be expanded into two power series such that

$$\left(\frac{p}{1-qs}\right)^{\eta/\lambda} = \sum_{k=0}^{\infty}\left[p^{\eta/\lambda}\binom{k + \frac{\eta}{\lambda} - 1}{k}q^k\right]s^k \tag{59}$$

and

$$\left[\frac{1 + \left(\frac{q}{p} - \gamma\right)(1-s)}{1 + \frac{q}{p}(1-s)}\right]^{k_0} = \sum_{k=0}^{\infty}\left[\sum_{j=0}^{\min[k_0,k]}\binom{k_0}{j}\binom{k_0 + k-j-1}{k-j}(1-\gamma)^{k_0-j}(pq)^{k-j}(\gamma-q)^j\right]s^k . \tag{60}$$

The expression in (59) is the probability generating function of a negative binomial random variable and the expression in (60) is the probability generating function of some other unnamed random variable. Therefore, $Pr[X(t)=k]$ is the convolution of the two sequences inside the brackets of expressions (59) and (60) (7).

The last stochastic model of migration that we shall consider is a linear growth, nonhomogeneous in time, birth-death-migration, compartmental process. This is by far the most realistic process that we will have considered here and is the most directly applicable to the situation of studying migration in the census tracts of the San Francisco Bay Area.

Census tracts and the area outside of the San Francisco/Oakland Standard Metropolitan Statistical Area are denoted by S_α, $\alpha=1,\ldots,s$. For a time interval (τ,t), let the populations of the tracts at the initial time τ be represented by a constant column vector:

$$x_\tau = \begin{pmatrix} x_{1\tau} \\ \vdots \\ x_{s\tau} \end{pmatrix} \tag{61}$$

and at time t by a random vector

$$X_t = \begin{pmatrix} X_{1t} \\ \vdots \\ X_{st} \end{pmatrix} \quad . \tag{62}$$

The value that X_t assumes is represented by

$$x_t = \begin{pmatrix} x_{1t} \\ \vdots \\ x_{st} \end{pmatrix} \tag{63}$$

where the $x_{\alpha t}$ are nonnegative integers.

For each w, $\tau < w < t$, a change in the population size of each tract S_α in the time interval $(w,w+\Delta)$ is assumed to take place according to the instantaneous probabilities:

$$\lambda_\alpha(t)x_{\alpha w}\Delta + o(\Delta) = \Pr\{\text{the size of tract } S_\alpha \text{ will increase by one in the interval } (w,w+\Delta)\}$$

$$\nu_{\alpha\beta}(t)x_{\alpha w}\Delta + o(\Delta) = \Pr\{\text{one individual will move from one tract } S_\alpha \text{ to } S_\beta \text{ in } (w,w+\Delta),\ \beta\neq\alpha\}$$

$$\mu_\alpha(t)x_{\alpha w}\Delta + o(\Delta) = \Pr\{\text{the size of tract } S_\alpha \text{ will decrease by one in the interval } (w,w+\Delta)\} .$$

We let

$$\nu_{\alpha\alpha}(t)x_w = -\Big[\sum_{\beta\neq\alpha} \nu_{\alpha\beta}(t)x_{\alpha w} + \mu_\alpha(t)x_{\alpha w}\Big]$$

and introduce the column vector:

$$\delta_\alpha = \begin{pmatrix} \delta_{1\alpha} \\ \vdots \\ \delta_{s\alpha} \end{pmatrix} \tag{64}$$

where the components are Kronecker deltas*.

Define $P(x_\tau, x_t)$ as the probability of having x_τ as the population at time t. The instantaneous probabilities satisfy the following system of differential equations:

$$\begin{aligned} \frac{\partial}{\partial t} P(x_\tau, x_t) &= -P(x_\tau, x_t) \sum_{\alpha=1}^{s} x_{\alpha t}\left[\lambda_\alpha(t) - \nu_{\alpha\alpha}(t)\right] \\ &+ \sum_{\alpha=1}^{s} P(x_\tau, x_t+\delta_\alpha)\lambda_\alpha(t)(x_{\alpha t}-1) \\ &+ \sum_{\substack{\alpha=1 \\ \alpha\neq\beta}}^{s}\sum_{\beta=1}^{s} P(x_\tau, x_t+\delta_\alpha-\delta_\beta)\nu_{\alpha\beta}(t)(x_{\alpha t}+1) \\ &+ \sum_{\alpha=1}^{s} P(x_\tau, x_t+\delta_\alpha)\mu_\alpha(t)(x_{\alpha t}+1) \end{aligned} \tag{65}$$

with initial conditions at $t=\tau$ of:

$$P(x_\tau, x_\tau) = 1, \quad P(x_\tau, x_t) = 0, \quad \text{for } x_t \neq x_\tau \ . \tag{66}$$

Direct computation shows that the corresponding generating functions satisfies the partial differential equation:

$$\begin{aligned} \frac{\partial}{\partial t} g_{X_t}(\underline{u};t) &= -\sum_{\alpha=1}^{s} \lambda_\alpha(t) u_\alpha(1-u_\alpha) \frac{\partial}{\partial u_\alpha} g_{X_t}(\underline{u};t) \\ &+ \sum_{\alpha=1}^{s} u_\alpha \nu_{\alpha\alpha}(t) \frac{\partial}{\partial u_\alpha} g_{X_t}(\underline{u};t) + \sum_{\substack{\alpha=1 \\ \alpha\neq\beta}}^{s}\sum_{\beta=1}^{s} u_\beta(t)\, \nu_{\alpha\beta}(t) \frac{\partial}{\partial u_\alpha} g_{X_t}(\underline{u};t) \\ &+ \sum_{\alpha=1}^{s} u_\alpha(t) \frac{\partial}{\partial u_\alpha} g_{X_t}(\underline{u};t) \ . \end{aligned} \tag{67}$$

*The Kronecker delta is defined as: $\delta_{ii}=1$ and $\delta_{ij}=0$, $j\neq i$.

Making the substitutions of $\mu_\alpha(t) = -\sum_{\beta=1}^{s} \nu_{\alpha\beta}(t)$ and $z_\alpha = 1-\mu_\alpha$ yields

$$\frac{\partial}{\partial t} g_{X_t}(\underline{z};t) = \sum_{\alpha=1}^{s} \left[\lambda_\alpha(t) z_\alpha(1-z_\alpha) + \sum_{\beta=1}^{s} \nu_{\alpha\beta}(t) z_\beta \right] \frac{\partial}{\partial z} g_{X_t}(\underline{z};t) \ . \qquad (68)$$

A simple explicit solution for the above equation is unknown at the present time. Therefore, although we have developed a realistic model of a migration we are unable to solve this process explicitly (6).

SECTION 6

STOCHASTIC MODELLING OF POPULATION GROWTH

Model Including In-migration Only

If it is assumed that the instantaneous probability of in-migration is not dependent on time or on the number of in-migrations in an area, then the Poisson process results. Thus, as presented in the introduction, the probability that there will be k in-migrations into an area by time t is:

$$p_k(t) = \frac{e^{-\lambda t}(\lambda t)^k}{k!} \qquad k=0,1,\ldots \ .$$

If this process is extended by allowing the parameter λ to be a function of time, the time dependent Poisson process results:

$$p_k(t) = \frac{\exp\{-\int_0^t \lambda(\tau)\,d\tau\}\left[\int_0^t \lambda(\tau)\,d\tau\right]^k}{k!} \qquad k=0,1,\ldots \ .$$

This process is made specific for the census tracts of the San Francisco-Oakland Standard Metropolitan Statistical Area by subscripting the process for the tracts. This creates the system of processes:

$$p_{k_i}(t) = \frac{\exp\{-\int_0^t \lambda_i(\tau)\,d\tau\}\left[\int_0^t \lambda_i(\tau)\,d\tau\right]^{k_i}}{k_i!} \ , \quad \begin{array}{l} k=0,1,\ldots; \\ k=1,\ldots,722 \text{ census tracts} \ . \end{array}$$

If $\lambda_i(t)$ is allowed to depend on the number of individuals in census tract i in such a way that the growth in an area is linear, the process becomes:

$$p_{k_i}(t) = \binom{k_i-1}{k_i-k_{i0}} e^{-k_{i0}\int_0^t \lambda_i(\tau)\,d\tau}\left[1 - e^{-\int_0^t \lambda_i(\tau)\,d\tau}\right]^{k_i-k_{i0}} \ ,$$

where k_{i0} is the number of individuals in tract i at time 0. This is known as the time-dependent Yule process.

For a process in which the instantaneous probability of in-migration is dependent on time and dependent on k_i in a non-linear way, the probability of k_i individuals being in census tract i at time t is:

$$p_{k_i}(t) = (-1)^{k_i-k_{i0}} \lambda_{k_{i0}}(t) \cdots \lambda_{k_i-1}(t) \left[\sum_{i=k_{i0}}^{k_i} \frac{e^{-\int_0^t \lambda_i(\tau)d\tau}}{\prod_{\substack{j=k_{i0} \\ j\neq i}}^{k} (\lambda_i-\lambda_j)} \right] ,$$

for $k_i > k_{i0}$.

Model Including In- and Out-Migration

A process can be constructed to allow for the possibilities of both in- and out-migration by including the instantaneous probabilities $\lambda_{k_i}(t)\Delta + o(\Delta)$ for in-migration and $\mu_{k_i}(t)\Delta+o(\Delta)$ for out-migration. For linear growth ($\lambda_{k_i}(t) \to k_i\lambda(t)$ and $\mu_{k_i}(t) \to k_i\mu(t)$), this process becomes

$$p_{k_i}(t) = \sum_{j=0}^{\min[k_{i0},k_i]} \binom{k_{i0}}{j} \binom{k_{i0}+k_i-j-1}{k_i-j} [\alpha(t)]^{k_{i0}-j} [\beta(t)]^{k_i-j} [1-\alpha(t)-\beta(t)]^j$$

for $k_i \geq 1$, where

$$p_0(t) = [\alpha(t)]^{k_{i0}}, \quad \alpha(t) = 1 - \frac{1}{e^{\gamma(t)} + \int_0^t \lambda(\tau)e^{\gamma(\tau)}d\tau} ,$$

$$\beta(t) = 1-e^{\gamma(t)}[1-\alpha(t)] , \quad \text{and } \gamma(t) = -\int_0^t [\lambda(\tau)-\mu(\tau)]d\tau .$$

Model Including In-, Out-, and Within-Migration

The model including in-, out-, and within-migration is based on parameter $\lambda_{k_i}(t)$ for in-migration, $\mu_{k_i}(t)$ for out-migration and $\gamma_{hjk_i}(t)$ for within-migration (from census tract h to census tract j). Parameter γ_{hjk_i} represents the movement of an individual from census tract h to census tract j. Therefore, from the standpoint of a particular census tract, say 1, $\gamma_{hlk_i}(t)$ represents the probability of in-migration from tract h to tract 1 and $\gamma_{ljk_i}(t)$ represents the probability of out-migration from tract 1 to tract j. For our purposes $\lambda_{k_i}(t)$ represents in-migration from outside the San Francisco-Oakland Standard Metropolitan Statistical Area and $\mu_{k_i}(t)$ represents out-migration to the area outside of the San Francisco-Oakland SMSA. Thus if our focus of attention is census tract 1, then the total instantaneous probability of in-migration of an individual is

$$\lambda_{k_i}(t)\Delta + \sum_{h=2}^{722} \gamma_{hlk_i}(t)\Delta + o(\Delta)$$

and the total instantaneous probability of out-migration of an individual is

$$\mu_{k_i}(t)\Delta + \sum_{j=2}^{722} \gamma_{ljk_i}(t)\Delta + o(\Delta) \quad .$$

These probabilities are similar to those in the in- and out-migration model except that we replace the $\lambda_{k_i}(t)$ of the in- and out-migration model with

$$\lambda_{k_i}(t) + \sum_{h=1}^{722} \lambda_{hik_i}(t)$$

and $\mu_{k_i}(t)$ with

$$\mu_{k_i}(t) + \sum_{j=1}^{722} \gamma_{ijk_i}(t) \quad .$$

With linear growth, we get:

$$p_{k_i}(t) = \sum_{j=0}^{\min[k_{i0},k_i]} \binom{k_{i0}}{j} \binom{k_{i0}+k_i-j-1}{k_i-j} [\alpha(t)]^{k_{i0}-j} [\beta(t)]^{k_i-j} [1-\alpha(t)-\beta(t)]$$

for $k_i \geq 1$, where

$$\alpha(t) = 1 - \frac{1}{\{e^{\gamma(t)} + \int_0^t [\lambda(t) + \sum_{h=1}^{722} \gamma_{hi}(\tau)] e^{\gamma(\tau)} d\tau\}} ,$$

$$\beta(t) = 1 - e^{\gamma(t)} [1-\alpha(t)] ,$$

and

$$\gamma(t) = -\int_0^t \{\lambda(\tau) + \sum_{h=1}^{722} \gamma_{hi}(\tau) - \mu(\tau) - \sum_{j=1}^{722} \gamma_{ij}(\tau)\} d\tau .$$

Birth and Death Complications in the Models

The inclusion of births and deaths into the above model presents little problem since they are analogous to in- and out-migrations. If $\lambda_1(t)\Delta+o(\Delta)$ and $\lambda_2(t)\Delta+o(\Delta)$ are the instantaneous probabilities of in-migration and birth, respectively, then $\lambda_3(t)\Delta+o(\Delta) = \lambda_1(t)\Delta+o(\Delta)+\lambda_2(t)\Delta+o(\Delta)$ is the instantaneous probability of an increase in population by either in-migration or birth. Also, if $\mu_1(t)\Delta+o(\Delta)$ and $\mu_2(t)\Delta+o(\Delta)$ are the instantaneous probabilities of out-migration and death, respectively, then $\mu_3(t)\Delta+o(\Delta) = \mu_1(t)\Delta + o(\Delta)+\mu_2(t)\Delta+o(\Delta)$ is the instantaneous probability of a decrease in population by either out-migration or death. The resulting process, with linear growth is:

$$p_{k_i}(t) = \sum_{j=0}^{\min[k_{i0},k_i]} \binom{k_{i0}}{j} \binom{k_{i0}+k_i-j-1}{k_{i-j}} [\alpha(t)]^{k_{i0}-j} [\beta(t)]^{k_i-j} [1-\alpha(t)-\beta(t)]^j$$

for $k_i \geq 1$, where

$$\alpha(t) = 1 - \frac{1}{\{e^{\gamma(t)} + \int_0^t \lambda_3(\tau) e^{\gamma(\tau)} d\tau\}} \quad ,$$

$$\beta(t) = 1 - e^{\gamma(t)}[1-\alpha(t)] \quad ,$$

and

$$\gamma(t) = -\int_0^t [\lambda_3(\tau) - \mu_3(\tau)] d\tau \quad .$$

SECTION 7

CENSUS MIGRATION DATA

Part of the 1970 Census of Population and Housing taken by the Bureau of the Census under the United States Department of Commerce was information on census tracts. The census was conducted as of April 1, 1970. The 1970 Census of Population and Housing on census tracts contains 241 reports. Each of these reports relates to a particular standard metropolitan statistical area (SMSA), and in some cases also covers certain areas adjacent to the SMSA. Although there were 247 recognized SMSA's in 1970, including four in Puerto Rico, there were no reports for six of these SMSA's because they were newly designated on the basis of the 1970 census results and census tracts had not been established for them previously.

Of the 247 SMSA's in 1970, 231 were defined and named in the Bureau of the Budget publication, Standard Metropolitan Statistical Areas: 1967, United States Government Printing Office, Washington, D.C. 20402. Since 1967, 16 SMSA's have been added, of which two were defined in January 1968 and an additional 14 were defined in February 1971 on the basis of the results of the 1970 census. Changes in SMSA boundaries or titles made after February 1971 were not recognized in the reports of the 1970 census. SMSA's are defined by the Office of Management and Budget (formerly Bureau of the Budget) (8).

Except in the New England states, a standard metropolitan statistical area is a county or group of contiguous counties which contains at least one city of 50,000 inhabitants or more, or "twin cities" with a combined population of at least 50,000. In addition to the county or counties containing such a city or cities, contiguous counties are included in an SMSA if, according to certain criteria, they are socially and economically integrated with the central city. In the New England States, SMSA's consist of towns and cities instead of counties. Each SMSA must include at least one central city, and the complete title of an SMSA identifies the central city or cities.

The concept of "Standard Metropolitan Statistical Areas" was developed to meet the need for the presentation of general-purpose statistics by agencies of the Federal Government, in accordance with specific criteria for defining such areas. On the basis of these criteria, "definitions" of the areas in terms of geographic boundaries are established by the Bureau of the Budget with the advice of the Federal Committee on Standard Metropolitan Statistical Areas, which is composed of representatives of the major Federal statistical agencies.

Standard definitions of metropolitan statistical areas were first issued in 1949 as "Standard Metropolitan Areas." They were developed to

replace four different sets of definitions then in use for various statistical series -- "metropolitan districts," "metropolitan counties," "industrial areas," and "labor market areas." Because of the use of these different definitions, it was not possible to relate the statistics on population, industrial production, labor force, and other series for the area in question since each series included a slightly different territory.

The primary objective in establishing standard definitions of metropolitan areas was thus to make it possible for all Federal statistical agencies to utilize the same boundaries in publishing statistical data useful for analysing metropolitan problems. The term "standard metropolitan area" has been changed to "standard metropolitan statistical area" in order to describe more accurately the objective of the definitions.

In order to achieve this purpose in a meaningful way, objective criteria of a quantitative character have been established to define the areas. The criteria were revised and reissued in March 1958, following review and minor revisions by the Federal Committee on Standard Metropolitan Statistical Areas.

The general concept of a metropolitan area is one of an integrated economic and social unit with a recognized large population nucleus. To serve the statistical purposes for which metropolitan areas are defined, their parts must themselves be areas for which statistics are usually or often collected. Thus, each standard metropolitan statistical area must contain one city of at least 50,000 inhabitants. The standard metropolitan statistical area will then include the county of such a central city, and adjacent counties that are found to be metropolitan in character and economically and socially integrated with the county of the central city. An mentioned above, in New England, the requirement with regard to a central city as a nucleus still holds, but the units comprising the area are the towns rather than counties. The county (or town in New England) is the basic statistical unit. A standard metropolitan statistical area may contain more than one city of 50,000 population. The largest city is considered the nucleus and usually gives the name to the area. The name may include other cities in the area. Standard metropolitan statistical areas may cross state lines.

Definitions, in terms of the counties or towns included, have been reviewed by the Federal Committee on Standard Metropolitan Statistical Areas on the basis of the criteria adopted in March 1958. One of the basic criteria for measuring economic integration, to determine whether additional counties should be included in an area definition, is the relationship of place of residence to place of work, involving outlying counties and the county of the central city. The volume of workers commuting was determined on the basis of data from the 1960 Census of Population, and the Bureau of Employment Security of the Department of Labor, in cooperation with its affiliated State Employment Security Agencies.

These definitions are designed to serve a wide variety of statistical and analytical purposes. Although they were developed for use by Federal agencies, they were expected to be useful to State and local governments, as well as private statistical agencies.

The definition of an individual standard metropolitan statistical area involved two considerations; first, a city or cities of specific population to constitute the central city and to identify the county in which it was

located as the central county; and, second, economic and social relationships with contiguous counties which were metropolitan in character, so that the periphery of the specific metropolitan area may be determined. Standard metropolitan statistical areas may cross state lines, if this is necessary, in order to include qualified contiguous counties.

In terms of population criteria, each standard metropolitan statistical area had to include at least: (a) one city with 50,000 or more inhabitants, or (b) two cities having contiguous boundaries and constituting, for general economic and social purposes, a single community with a combined population of at least 50,000, the smaller of which must have a population of at least 15,000. Also, if two or more adjacent counties each had a city of 50,000 inhabitants or more (twin cities) and the cities were within 20 miles of each other (city limits to city limits), they were included in the same area unless there is definite evidence that the two cities are not economically and socially integrated.

The criteria of metropolitan character relate primarily to the attributes of the county as a place of work or as a home for a concentration of nonagricultural workers. Specifically, these criteria are: (a) at least 75% of the labor force of the county must be in the nonagricultural labor force[1], and one of the following three conditions must be met, (b) it must have 50% or more of its population living in contiguous minor civil divisions[2] with a density of at least 150 persons per square mile, in an unbroken chain of minor civil divisions with such density radiating from a central city[3] in the area, or (c) the number of nonagricultural workers employed in the county must equal to least 10% of the number of nonagricultural workers employed in the county containing the largest city in the area, or be the place of employment of 10,000 nonagricultural workers, or (d) the nonagricultural labor force living in the county must equal to least 10% of the number of the nonagricultural labor force living in the county containing the largest city in the area, or be the place of residence of a nonagricultural labor force of 10,000.

The criteria of integration relate primarily to the extent of economic and social communication between the outlying counties and central county. A county is regarded as integrated with the county or counties containing

---

[1] Nonagricultural labor force is defined as those employed in nonagricultural occupations, those experienced unemployed whose last occupation was a nonagricultural occupation, members of the Armed Forces, and new workers.

[2] A contiguous minor civil division either adjoins a central city in a standard metropolitan statistical area or adjoins an intermediate minor civil division of qualifying population density. There is no limit to the number of tiers of contiguous minor civil divisions so long as the minimum density requirement is met in each tier.

[3] Central cities are those appearing in the standard metropolitan statistical area title.

the central cities of the area if either of the following criteria are met: (a) if 15% of the workers living in the county work in the county or counties containing central cities of the area, or (b) if 25% of those working in the county live in the county or counties containing central cities of the area.

The following guidelines were used in determining titles for standard metropolitan statistical areas: (a) the name of the area is that of the largest city, or (b) the addition of up to two city names may be made in the area title, on the basis and in the order of the following criteria: (1) the additional city or cities have at least 250,000 inhabitants, and (2) the additional city or cities have a population of one-third or more of that of the largest city and a minimum population of 25,000, except that both city names are used in those instances where cities qualify under the criterion describing a standard metropolitan statistical area containing two cities. Also, in addition to city names, the area titles will contain the name of the state or states included in the area.

The definitions and titles of standard metropolitan statistical areas are established by the Bureau of the Budget with the advice of the Federal Committee on the Standard Metropolitan Statistical Areas. This Committee is composed of representatives of the major statistical agencies of the Federal Government. In applying the foregoing criteria, data from the following sources are used by the Committee: Population, labor force, density, and occupational data -- Bureau of the Census and Bureau of Employment Security. Employment by place of work -- Bureau of Old-Age and Survivors Insurance, Department of Labor, Department of Defense, and Civil Service Commission. Volume of commuting -- Bureau of the Census and Bureau of Employment Security (8).

Within these SMSA's, census tracts are defined. Census tracts are small areas into which large cities and adjacent areas have been divided for statistical purposes. Tract boundaries were established cooperatively by a local committee and Bureau of the Census. Tracts were generally designed to be relatively uniform with respect to population characteristics, economic status, and living conditions. The average tract has about 4,000 residents. Tract boundaries are established with the intention of being maintained over a long time so that comparisons may be made from census to census.

In the decennial censuses, the Bureau of the Census tabulated population and housing information for each census tract. The practice of local agencies tabulating locally collected data by tracts has increased the value of census tract data in many areas.

Tracts are generally numbered in a consecutive series, with separate series for the central city and for each county. Insofar as possible, the numbers are consecutive within each city, community township, and the like. Tracts which contain no population and no housing units are not listed in published reports of the Bureau of the Census. Tracts populated only by crews of vessels are identified by the tract number suffix "99."

The Bureau of the Census, at the request of the late Dr. Walter Laidaw,

tabulated census tract data from the 1910 census for New York and seven other cities with a population of over 500,000. Tract data were again tabulated for the same eight cities in 1920, and in 1930 this number was increased to 18. In 1940, tract data were tabulated for 60 cities, some with adjacent tracted areas; and, beginning in 1940, housing data were added to the population data in the tract reports. In 1950, final reports were published for 64 tracted areas, many of which included statistics for two or more large cities. By 1960, the program had expanded to include reports for 180 tracted areas (of which three were in Puerto Rico).

Tract statistics from the 1970 census are published for 241 areas, 238 in the United States and three in the Commonwealth of Puerto Rico. All of these tracted areas are standard metropolitan statistical areas (SMSA's). The goal of extending the census tract program to all SMSA's was achieved in the 1970 census except for six areas that were designated as SMSA's on the basis of the 1970 census count and had not been tracted.

Much of the credit for the growing interest in tract data belongs to the late Howard Whipple Green of Cleveland. He aroused the interest of research workers in numerous cities in the potential usefulness of tract statistics for the analysis of sociological, marketing, and administrative problems. In his capacity as Chairman of the Committee on Census Enumeration Areas of the American Statistical Association for 25 years, he accepted the responsibility for appointing a Census Tract Key Person in each area where tracts were established, for providing guidance on delineating and maintaining census tracts, and for maintaining a census tract library. These duties were assumed by the Bureau of the Census in 1955. However, the Census Bureau no longer appoints Census Tract Key Persons. They are selected by the local census tract committees.

One of the principles followed in relation to census tracts is to preserve comparability from census to census. Keeping tract boundaries unchanged makes possible the study of changes in social and economic characteristics of neighborhoods. There are, nonetheless, several situations where boundaries of individual tracts are changed. For example, it sometimes is necessary to change the boundaries of tracts to add small areas annexed to a city. Similarly, changes in tract boundaries occur when territory is detached from a city or separately incorporated. Changes may also occur in physical features that are used as tract boundaries, such as street or highway relocations. The census tract limits are changed to conform with the revised feature or to follow another nearby visible feature. Census tracts with very large increases in population are subdivided into two or more smaller tracts. On the other hand, a reexamination of the existing tract boundaries may result in modifications of boundaries to provide larger or more homogeneous units.

Under the reports of the Bureau of the Census there are two variables that relate to the subject of the migration or population mobility. Under the population portion of the reports there is an item entitled "Residence in 1965." Residence on April 1, 1965 is the usual place of residence five years before enumeration. The category "same house" includes all persons five years old and over who did not move during the five years as well as

those who had moved but by 1970 had returned to their 1965 residence. The category "different house" includes persons who, on April 1, 1965, lived in the United States in a different house from the one they occupied on April 1, 1970, and for whom sufficient information concerning the 1965 residence was collected. These persons were subdivided into three groups according to their 1965 residence in or outside a standard metropolitan statistical area: "in central city of this SMSA," "in other parts of this SMSA," and "outside this SMSA." The category "abroad" includes those with residence in a foreign country or outlying area of the United States in 1965.

Under the Housing Characteristics portion of the 1970 Bureau of the Census reports there is a variable entitled "Year moved into unit." Data on year moved into unit are based on the information reported for the head of the household. The question refers to the year of the latest move. Thus, if the head moved back into a unit he had previously occupied or if he moved from one apartment to another in the same building, the year he moved into his present unit was to be reported. "Year moved into unit" was reported in five categories: 1968 to March 1970, 1965 to 1967, 1960 to 1964, 1950 to 1959, and 1949 or earlier (2).

SECTION 8

ESTIMATION OF MIGRATION BY CENSUS MIGRATION DATA

In this section a procedure is developed for estimation of migrants based on the 1970 census item "Year Moved Into Unit" found in the Housing Characteristics portion of the 1970 Census of Population and Housing. The exact source is Table H-2 of the Census Bureau's report on census tracts which is Series PHC (Population and Housing Characteristics) (1) of the Joint Population-Housing Reports entitled "Census Tract Reports." This item is found in the Second Count of the Census Bureau's Computer Summary Tapes. The item "Year Moved Into Unit" is based on a sample rate of 15%.

The variable "Year Moved Into Unit" (YMIU) provides information on the number of occupied housing units in census tracts and the following data on year moved into unit based on the information reported for the head of household: 1968 - March 1970, 1965 - 1967, 1960-1964, 1950-1959, 1949 or earlier.

The question refers to the year of latest move. Thus, if the head moved back into a unit he had previously occupied or if he moved from one apartment to another in the same building, the year he moved into his present unit was to be reported. Since the above five categories involve intervals in time it is necessary to choose a time within each of these intervals that will be used in the estimation procedure. The midpoint in each of the first four categories is the obvious choice. The fifth category (1949 or earlier) presents a problem since it not an interval. For convenience the date 1/1/49 will be used to represent the 1949 or earlier category. The midpoints are as follows:

Year Moved Into Unit Categories	Midpoints
1968 - March 1970	2/15/69
1965 - 1967	7/1/66
1960 - 1964	7/1/62
1950 - 1959	1/1/55
1949 or earlier	1/1/49

These midpoints assume that the heads of households moved into their respective units uniformly throughout the year. We, therefore, compute the amount of time between these midpoints and the midpoint of the study period which is 7/1/70 (the midpoint for the year 1969-1971). These lengths of time are presented in the following table:

Midpoints of YMIU Categories	-	Midpoint of Study Period	Length of Interval (in years)
2/15/69	-	7/1/70	1.375 = c_1
7/1/66	-	7/1/70	4.0 = c_2
7/1/62	-	7/1/70	8.0 = c_3
1/1/55	-	7/1/70	15.5 = c_4
1/1/49	-	7/1/70	21.5 = c_5

Next it is assumed that the proportion of all occupied units in each YMIU category represents the proportion of population in a census tract in that category:

YMIU Categories	Proportion of All Occupied Units
1968 - March 1970	p_1
1965 - 1967	p_2
1960 - 1964	p_3
1950 - 1959	p_4
1949 or earlier	p_5

Another assumption we can make is that all moves represent in-migration from outside the SMSA. This assumption certainly is not a very accurate portrayal of the true state of affairs but it does simplify matters and it is likely that in a relative sense the assumption works pretty well. That is to say, if two census tracts have a large difference in the proportion of all occupied units having heads of households arriving in the YMIU category of 1940 or earlier, then it is more likely that this represents a difference in number of in-migrants as well as movers within their respective census tracts. An alternative assumption would be that a certain percentage, say 50%, of the proportions in each YMIU category represent in-migrants from outside the SMSA.

Whatever assumption about the amount of the proportions in each YMIU category representing in-migrants from outside the SMSA is adopted, it is further assumed that the age distribution of these in-migrants follows the age distribution of the census tract in question. A more simplifying assumption would be that the age distribution of in-migrants follows the age distribution of the entire SMSA.

For each census tract $\sum_{i=1}^{5} c_i p_i$ is computed to yield the weighted average length of exposure (the weights being the proportion of all occupied units in each length of interval category). This weighted average is then multiplied by the number of white males and white females in each census tract resulting in estimates of length of exposure given the aforementioned assumptions.

TYPING GUIDE SHEET

SECTION 9

VARIABLES IN ASBESTOS-CANCER DATA ANALYSIS

The variables entered into the data analysis for each census tract were chrysotile asbestos fibers per liter of drinking water, observed to expected ratios of cancer incidence cases (standard morbidity ratios) by body site and sex for the white population for the years 1969-1971, socioeconomic status measured by median family income and median school years completed, proportions of census tract populations that were single, separated, widowed, or divorced, proportions of census tract populations that were employed in industries that might have exposure to asbestos, and years of exposure. See Table 2 for a list of cancer body sites and cancer body site groupings.

Asbestos data were obtained through the analysis of 353 water samples. Observed cancer incidence cases were taken from the Third National Cancer Survey and expected cases were calculated using Census population data and observed cases in the indirect standardization procedure. Observed cancer incidence rates for the entire SMSA were used as the standard in this procedure. Years of exposure were estimated by the aforementioned method. The other variables were obtained from the 1970 Census of Population and Housing. A more detailed description of the above variables and their sources is included in the U.S. Environmental Protection Agency's report entitled Asbestos in Domestic Water Supplies in Five California Counties.

TABLE 2

CANCER SITES ANALYZED

No.	Name
1	All sites
2	All digestive
3	Digestive tract
4	Esophagus
5	Stomach
6	Small intestine
7	Colon
8	Rectum
9	Digestive related organs
10	Liver
11	Gall bladder
12	Pancreas
13	Peritoneum
14	Digestive organ NOS
15	All respiratory
16	Larynx
17	Trachea, bronchus and lung
18	Pleura
19	Mediastinum
20	Breast
21	Female reproductive
22	Cervix uteri
23	Corpus uteri
24	Ovary
25	Prostate
26	Urinary
27	Kidney
28	Bladder
29	Brain
30	Thyroid
31	Hodgkin's disease
32	Leukemia
33	Lung small cell CA
34	Lung squamous
35	Lung adenoca NOS

SECTION 10

RESULTS

Descriptive statistics were calculated for all variables and a computed variable entitled "asbestos x exposure." Asbestos x exposure was computed for males and females by multiplying asbestos values by sex-specific years of exposure for each census tract. Table 3 presents means, standard deviations, and the number of valid and missing cases (census tracts) for all variables. Of interest in this table is the fact that the majority of mean observed to expected ratios of cancer incidence for both males and females are greater than one indicating that more cancer cases were observed than were expected based on the age distributions of the census tracts studied.

Pearson product moment correlation coefficients between cancer incidence ratios and asbestos by cancer body site and sex across census tracts are presented in Table 4. The significance of each correlation is also presented. Significance is determined by the t-test for the null hypothesis that the population correlation coefficient was equal to zero. For males the most significant body sites were 4 esophagus, 5 stomach, 13 peritoneum, 15 all respiratory, 17 trachea-lung, and 28 bladder. For females the most significant sites were 2 all digestive, 4 esophagus, 5 stomach, 9 digestive related organs, 11 gall bladder, 12 pancreas, and 13 perioteneum. Consistency between the sexes on this significance is observed for sites 4 esophagus, 5 stomach, and 13 peritoneum. All of the above correlations indicate positive relationships between asbestos and cancer. Seven negative correlation coefficients were observed but none of these were found to be significant.

Table 5 displays Pearson product moment correlation coefficients between cancer incidence ratios and estimated years of exposure by site and sex across census tracts. All but a few coefficients are statistically significantly different from zero. The significant coefficients were all positive indicating positive relationships between exposure and cancer. There was a negative coefficient but it was not significant. The significance for these correlations was much greater than for the correlations of cancer with asbestos.

Multiple regression procedures were used to examine the relationship between census tract asbestos fiber values and cancer incidence while taking into account other variables such as socio-economic status, asbestos industry, and years of exposure. To make these rates comparable to those of the original asbestos-cancer study, years of exposure were entered into the regressions as an independent variable. Since a natural logarithm transformation of the dependent variable (observed to expected ratio of cancer incidence) was executed in the original study, the identical transformation was used here. To avoid taking the natural logarithm of zero (which results

in negative infinity), a small constant (.01) was added to each observed to expected ratio prior to the transformation by natural logarithm. Zero cancer cases were observed in some census tracts for the more rare body sites such as pleura and peritoneum.

The log linear regressions were performed using the multiple regression program of the Statistical Package for the Social Sciences (SPSS) (9). The regressions were run for white males and white females, using what was termed in the original study, the "Fine" asbestos assignments. The "Fine" asbestos assignments were distinguished from the "Gross" assignments. "Fine" assignments were made by applying asbestos sample values of the 353 water samples to individual census tracts whereever possible. "Gross" assignments were made by using the asbestos values from the 353 was samples to characterize water district asbestos levels and applying those values to census tracts within their respective water districts. Only the "Fine" asbestos assignments were used in the present analysis.

Regression coefficients and their significance are presented in Table 6. Since a large number of coefficients were tested in these regressions only those with significance less than .01 will be considered statistically significant. For asbestos in the male sample, body cancer sites 13 peritoneum, 15 all respiratory, and 17 trachea-lung were significant. For asbestos in the female sample, body cancer sites 2 all digestive, 5 stomach, 9 digestive related organs, 11 gall bladder, 12 pancreas, and 20 breast were significant. Although a number of negative regression coefficients for asbestos were observed for both sexes, none were found to be signficant. For years of exposure in both sexes nearly all body sites of cancer were significant. None of the exposure regression coefficients were negative.

Two other regression models were employed. One used asbestos, years of exposure, and an interaction variable (asbestos x years of exposure) and one omitted asbestos and years of exposure but used the asbestos x years of exposure interaction term. The interaction term indicates the total dose of asbestos. The former model was executed but will not be reported upon here due to problems with multicollinearity between asbestos, exposure and the interaction term. Table 7 shows the regression coefficients and their significance for the model with interaction terms and covariables only (model does not include the individual terms of asbestos and exposure). For both sexes there were many positive regression coefficients that were significantly different from zero. None of the negative coefficients, of which there were three, were signficantly different from zero. Table 8 lists the significant cancer body sites ($p<.01$) by sex. These results show that positive relationships were observed between asbestos x years of exposure and cancer for digestive organs, digestive related organs, respiratory organs, and urinary organs. In general these were highly significant results.

TABLE 3

DESCRIPTIVE STATISTICS FOR ANALYZED VARIABLES

	VARIABLE	MEAN	STANDARD DEVIATION	VALID CASES	MISSING CASES
	Asbestos	5.99×10^6	9.01×10^6	705	16
	Proportion in Asbestos Industry	.049	.023	721	0
	Median Years Schooling	12.85	5.80	721	0
	Median Family Income	11915.3	3978.8	710	11
	Proportion Unmarried	.417	.135	721	0
	Male Years of Exposure	12490.0	8475.8	721	0
	Observed to Expected Cancer Ratios: Male				
1	All Sites	1.273	5.153	721	0
2	All Digestive	1.052	1.510	721	0
3	Digestive Tract	1.032	1.635	721	0
4	Esophagus	1.211	4.465	721	0
5	Stomach	1.023	2.792	721	0
6	Small Intestine	1.092	7.759	721	0
7	Colon	.988	1.779	721	0
8	Rectum	1.071	2.224	721	0
9	Digestive Related Organs	1.134	2.692	721	0
10	Liver	.957	2.854	721	0
11	Gall Bladder	1.326	7.821	721	0
12	Pancreas	1.163	3.745	721	0
13	Peritoneum	.876	3.980	721	0
14	Digestive Organ NOS	.893	4.251	721	0
15	All Respiratory	2.360	31.798	721	0
16	Larynx	2.487	37.829	721	0
17	Trachea/Bronchus/Lung	2.474	37.019	721	0
18	Pleura	1.008	6.561	721	0
19	Mediastinum	.981	7.701	721	0
20	Breast	.905	5.148	721	0
21	Female Reproductive	-	-	-	-
22	Cervix Uteri	-	-	-	-
23	Corpus Uteri	-	-	-	-
24	Ovary	-	-	-	-
25	Male Reproductive & Prostate	1.149	2.307	721	0

TABLE 3 (continued)

DESCRIPTIVE STATISTICS FOR ANALYZED VARIABLES

VARIABLE		MEAN	STANDARD DEVIATION	VALID CASES	MISSING CASES
Observed to Expected Cancer Ratios: Male					
26	Urinary	1.035	1.618	721	0
27	Kidney	1.028	2.367	721	0
28	Bladder	1.040	2.122	721	0
29	Brain	1.032	2.758	721	0
30	Thyroid	.935	2.781	721	0
31	Hodgkin's Disease	1.459	11.720	721	0
32	Leukemia	1.099	3.279	721	0
33	Lung Small Cell CA	1.111	3.206	721	0
34	Lung Squamous	1.144	2.527	721	0
35	Lung Adenoca NOS	1.035	3.473	721	0
Female Years of Exposure		13324.3	9286.9		
Observed to Expected Cancer Ratios: Female					
1	All Sites	1.323	4.317	721	0
2	All Digestive	1.402	5.021	721	0
3	Digestive Tract	1.448	5.957	721	0
4	Esophagus	1.271	7.343	721	0
5	Stomach	1.290	5.844	721	0
6	Small Intestine	.903	4.812	721	0
7	Colon	1.555	7.729	721	0
8	Rectum	1.352	6.212	721	0
9	Digestive Related Organs	1.238	3.213	721	0
10	Liver	1.736	12.854	721	0
11	Gall Bladder	1.172	4.138	721	0
12	Pancreas	1.165	3.183	721	0
13	Peritoneum	.979	6.022	721	0
14	Digestive Organ NOS	.987	5.068	721	0
15	All Respiratory	1.539	8.097	721	0

TABLE 3 (continued)

DESCRIPTIVE STATISTICS FOR ANALYZED VARIABLES

VARIABLE		MEAN	STANDARD DEVIATION	VALID CASES	MISSING CASES
Observed to Expected Cancer Ratios: Female					
16	Larynx	1.159	11.527	721	0
17	Trachea/Bronchus/Lung	1.554	8.296	721	0
18	Pleura	.851	12.764	721	0
19	Mediastinum	.756	6.416	721	0
20	Breast	1.083	1.860	721	0
21	Female Reproductive	1.450	6.063	721	0
22	Cervix Uteri	2.095	17.240	721	0
23	Corpus Uteri	1.169	4.178	721	0
24	Ovary	1.683	10.345	721	0
25	Male Reproductive & Prostate	-	-	-	-
26	Urinary	1.344	5.842	721	0
27	Kidney	1.694	14.039	721	0
28	Bladder	1.131	3.548	721	0
29	Brain	.965	2.414	721	0
30	Thyroid	1.228	7.768	721	0
31	Hodgkin's Disease	3.045	49.749	721	0
32	Leukemia	2.063	24.319	721	0
33	Lung Small Cell CA	1.007	5.732	721	0
34	Lung Squamous	1.765	13.985	721	0
35	Lung Adenoca NOS	1.020	3.270	721	0
Male Years of Exposure x Asbestos		7.46×10^{10}	1.39×10^{11}	721	0
Female Years of Exposure x Asbestos		8.22×10^{10}	1.56×10^{11}	721	0

TABLE 4

PEARSON PRODUCT MOMENT CORRELATIONS AND SIGNIFICANCE* FOR OBSERVED TO EXPECTED CANCER RATIOS AND ASBESTOS BY SITE AND SEX

		MALE		FEMALE	
	SITE	r	SIGN.	r	SIGN.
1	All Sites	.097	.01	.084	.025
2	All Digestive	.069	.05	.104	.005
3	Digestive Tract	.075	.05	.053	N.S.
4	Esophagus	.110	.005	.121	.005
5	Stomach	.114	.005	.136	.005
6	Small Intestine	.041	N.S.	.018	N.S.
7	Colon	.022	N.S.	.060	N.S.
8	Rectum	.087	.025	.069	.05
9	Digestive Rel. Organs	.083	.025	.191	.0005
10	Liver	.034	N.S.	.033	N.S.
11	Gall Bladder	.027	N.S.	.171	.0005
12	Pancreas	.084	.025	.169	.0005
13	Peritoneum	.114	.005	.112	.005
14	Digestive Organs NOS	.035	N.S.	.001	N.S.
15	All Respiratory	.102	.005	.049	N.S.
16	Larynx	-.023	N.S.	-.020	N.S.
17	Trachea-Lung	.105	.005	.044	N.S.
18	Pleura	-.007	N.S.	.099	.01
19	Mediastinum	.031	N.S.	.001	N.S.
20	Breast	.004	N.S.	.095	.01
21	Female Reproductive	-	-	.052	N.S.
22	Cervix Uteri	-	-	-.037	N.S.
23	Corpus Uteri	-	-	-.011	N.S.
24	Ovary	-	-	.095	.01
25	Male Reproductive	.066	.05	-	-
26	Urinary	.064	N.S.	.065	N.S.
27	Kidney	-.004	N.S.	.096	.01
28	Bladder	.110	.005	.047	N.S.
29	Brain	.010	N.S.	.031	N.S.
30	Thyroid	.036	N.S.	.061	N.S.
31	Hodgkin's Disease	.028	N.S.	.021	N.S.
32	Leukemia	.048	N.S.	.095	.01
33	Lung Small Cell CA	.043	N.S.	-.020	N.S.
34	Lung Squamous	.088	.025	.080	.025
35	Lung Adenoca NOS	.088	.025	.056	N.S.

*Based on H_0: $\rho=0$, $n=694$

TABLE 5

PEARSON PRODUCT MOMENT CORRELATIONS AND SIGNIFICANCE* FOR OBSERVED TO EXPECTED CANCER RATIOS AND YEARS OF EXPOSURE BY SITE AND SEX

		MALE		FEMALE	
	SITE	r	SIGN.	r	SIGN.
1	All Sites	.150	.0005	.087	.025
2	All Digestive	.242	.0001	.222	.0001
3	Digestive Tract	.269	.0001	.259	.0001
4	Esophagus	.141	.0005	.078	.025
5	Stomach	.226	.0001	.148	.0005
6	Small Intestine	.075	.05	.136	.0005
7	Colon	.324	.0001	.299	.0001
8	Rectum	.167	.0005	.257	.0001
9	Digestive Rel. Organs	.239	.0001	.197	.0005
10	Liver	.149	.0005	.114	.005
11	Gall Bladder	.136	.0005	.145	.0005
12	Pancreas	.243	.0001	.190	.0005
13	Peritoneum	.116	.005	.116	.005
14	Digestive Organs NOS	.109	.005	.099	.01
15	All Respiratory	.255	.0001	.277	.0001
16	Larynx	.081	.025	-.013	N.S.
17	Trachea-Lung	.252	.0001	.276	.0001
18	Pleura	.091	.025	.036	N.S.
19	Mediastinum	.041	N.S.	.119	.005
20	Breast	.060	N.S.	.260	.0001
21	Female Reproductive	-	-	.217	.0001
22	Cervix Uteri	-	-	.080	.025
23	Corpus Uteri	-	-	.337	.0001
24	Ovary	-	-	.253	.0001
25	Male Reproductive	.175	.0005	-	-
26	Urinary	.272	.0001	.261	.0001
27	Kidney	.242	.0001	.169	.0005
28	Bladder	.258	.0001	.232	.0001
29	Brain	.206	.0001	.207	.0001
30	Thyroid	.151	.0005	.199	.0005
31	Hodgkin's Disease	.191	.0005	.156	.0005
32	Leukemia	.234	.0001	.191	.0005
33	Lung Small Cell CA	.129	.005	.153	.0005
34	Lung Squamous	.188	.0005	.128	.005
35	Lung Adenoca NOS	.261	.0001	.201	.0001

*Based on H_0: $\rho=0$, $n=694$

TABLE 6

REGRESSION COEFFICIENTS AND SIGNIFICANCE* FOR ASBESTOS AND YEARS OF EXPOSURE BY SITE AND SEX

		MALE		FEMALE	
	SITE	ASBESTOS	EXPOSURE	ASBESTOS	EXPOSURE
1	All Sites	.966E-08 .012	.153E-04 .0001	.918E-08 .021	.589E-05 .129
2	All Digestive	.847E-08 .191	.497E-04 .00001	.216E-07 .002	.408E-04 .00001
3	Digestive Tract	.127E-07 .087	.624E-04 .00001	.147E-07 .056	.539E-04 .00001
4	Esophagus	.210E-07 .057	.625E-04 .00001	.181E-07 .064	.299E-04 .002
5	Stomach	.237E-07 .038	.914E-04 .00001	.362E-07 .002	.559E-04 .0001
6	Small Intestine	.246E-08 .748	.125E-04 .128	.153E-08 .831	.215E-04 .002
7	Colon	-.889E-09 .927	.896E-04 .00001	.110E-07 .236	.801E-04 .00001
8	Rectum	.201E-07 .079	.656E-04 .00001	.135E-07 .237	.790E-04 .00001
9	Digestive Rel. Organs	.143E-07 .199	.820E-04 .00001	.459E-07 .0001	.586E-04 .00001
10	Liver	.374E-08 .714	.543E-04 .00001	-.347E-08 .714	.349E-04 .00001
11	Gall Bladder	-.382E-09 .968	.405E-04 .0001	.400E-07 .0001	.435E-04 .0001
12	Pancreas	.134E-07 .247	.885E-04 .00001	.398E-08 .001	.526E-04 .0001
13	Peritoneum	.231E-07 .003	.205E-04 .013	.179E-07 .013	.197E-04 .005
14	Digestive Org. NOS	.543E-08 .488	.281E-04 .001	-.327E-08 .700	.237E-04 .004
15	All Respiratory	.212E-07 .003	.494E-04 .00001	.691E-08 .500	.742E-04 .00001

TABLE 6 (continued)

REGRESSION COEFFICIENTS AND SIGNIFICANCE* FOR ASBESTOS AND YEARS OF EXPOSURE BY SITE AND SEX

		MALE		FEMALE	
	SITE	ASBESTOS	EXPOSURE	ASBESTOS	EXPOSURE
16	Larynx	-.776E-08 .375	.245E-04 .009	-.735E-09 .875	-.121E-05 .790
17	Trachea-Lung	.226E-07 .002	.498E-04 .00001	.548E-08 .596	.751E-04 .00001
18	Pleura	-.771E-09 .907	.144E-04 .042	.781E-08 .018	.328E-05 .307
19	Mediastinum	.308E-08 .578	.955E-05 .108	-.233E-08 .629	.147E-04 .002
20	Breast	-.172E-08 .801	.132E-04 .071	.177E-07 .003	.276E-04 .0001
21	Female Reproductive	-	-	.718E-08 .316	.371E-04 .00001
22	Cervix Uteri	-	-	-.771E-08 .525	.446E-04 .0001
23	Corpus Uteri	-	-	-.708E-08 .429	.649E-04 .00001
24	Ovary	-	-	.129E-07 .251	.771E-04 .00001
25	Male Reproductive	.981E-08 .249	.458E-04 .00001	-	-
26	Urinary	.143E-07 .145	.667E-04 .00001	.182E-07 .117	.764E-04 .00001
27	Kidney	-.201E-08 .857	.780E-04 .00001	.259E-07 .023	.483E-04 .0001
28	Bladder	.269E-07 .012	.704E-04 .00001	.539E-08 .642	.709E-04 .00001
29	Brain	-.224E-08 .845	.634E-04 .00001	.730E-08 .513	.490E-04 .0001
30	Thyroid	.564E-08 .589	.435E-04 .0001	.959E-08 .413	.549E-04 .0001

TABLE 6 (continued)

REGRESSION COEFFICIENTS AND SIGNIFICANCE* FOR ASBESTOS AND YEARS OF EXPOSURE BY SITE AND SEX

		MALE		FEMALE	
	SITE	ASBESTOS	EXPOSURE	ASBESTOS	EXPOSURE
31	Hodgkin's Disease	.825E-08 .468	.513E-04 .0001	-.292E-08 .784	.453E-04 .0001
32	Leukemia	.137E-07 .246	.791E-04 .00001	.169E-07 .149	.629E-04 .00001
33	Lung Small Cell CA	.226E-07 .055	.489E-04 .0001	-.532E-08 .594	.472E-04 .0001
34	Lung Squamous	.257E-09 .019	.648E-04 .00001	.192E-07 .094	.444E-04 .0001
35	Lung Adenoca NOS	.204E-07 .076	.806E-04 .00001	.119E-07 .297	.534E-04 .0001

*Based on H_0: B=0

TABLE 7

REGRESSION COEFFICIENTS AND SIGNIFICANCE FOR ASBESTOS X YEARS OF EXPOSURE BY SITE AND SEX

		MALE		FEMALE	
	SITE	B	SIGN.	B	SIGN.
1	All Sites	.594E-12	.013	.234E-12	.278
2	All Digestive	.144E-11	.011	.173E-11	.0001
3	Digestive Tract	.194E-11	.0001	.206E-11	.0001
4	Esophagus	.222E-11	.001	.154E-11	.005
5	Stomach	.374E-11	.0001	.312E-11	.0001
6	Small Intestine	.501E-12	.291	.732E-12	.064
7	Colon	.233E-11	.0001	.249E-11	.0001
8	Rectum	.289E-11	.0001	.277E-11	.0001
9	Digestive Rel. Organs	.316E-11	.0001	.321E-11	.0001
10	Liver	.173E-11	.007	.104E-11	.046
11	Gall Bladder	.167E-11	.005	.281E-11	.0001
12	Pancreas	.320E-11	.0001	.312E-11	.0001
13	Peritoneum	.235E-11	.0001	.151E-11	.0001
14	Digestive Organs NOS	.132E-11	.006	-.249E-13	.958
15	All Respiratory	.186E-11	.0001	.181E-11	.002
16	Larynx	-.145E-12	.788	-.126E-12	.622
17	Trachea-Lung	.193E-11	.0001	.173E-11	.004
18	Pleura	.485E-12	.233	.462E-12	.011
19	Mediastinum	.547E-12	.108	.513E-12	.053
20	Breast	.190E-12	.654	.110E-11	.001
21	Female Reproductive	-	-	.696E-11	.088
22	Cervix Uteri	-	-	.348E-12	.608
23	Corpus Uteri	-	-	.879E-12	.091
24	Ovary	-	-	.243E-11	.0001
25	Male Reproductive	.158E-11	.003	-	-
26	Urinary	.259E-11	.0001	.287E-11	.0001
27	Kidney	.193E-11	.009	.225E-11	.0001
28	Bladder	.368E-11	.0001	.231E-11	.0001
29	Brain	.199E-11	.006	.153E-11	.014
30	Thyroid	.119E-11	.067	.174E-11	.008
31	Hodgkin's Disease	.139E-11	.051	.113E-11	.059
32	Leukemia	.305E-11	.0001	.217E-11	.001
33	Lung Small Cell CA	.230E-11	.002	.554E-12	.319
34	Lung Squamous	.291E-11	.0001	.212E-11	.001
35	Lung Adenoca NOS	.341E-11	.0001	.149E-11	.020

TABLE 8

SIGNIFICANT CANCER BODY SITES (P<.01) FOR ASBESTOS X YEARS OF EXPOSURE BY SEX WITH SIGNIFICANCE

MALE			FEMALE		
	SITE	SIGN.		SITE	SIGN.
			2.	All Digestive	.0001
3.	Digestive Tract	.0001	3.	Digestive Tract	.0001
4.	Esophagus	.001	4.	Esophagus	.005
5.	tomach	.00001	5.	Stomach	.0001
7.	lon	.00001	7.	Colon	.0001
8.	Rectum	.00001	8.	Rectum	.0001
9.	Digestive Rel. Organs	.00001	9.	Digestive Rel. Organs	.0001
10.	Liver	.007			
11.	Gall Bladder	.005	11.	Gall Bladder	.0001
12.	Pancreas	.0001	12.	Pancreas	.0001
13.	Peritoneum	.0001	13.	Peritoneum	.0001
14.	Digestive Organs NOS	.006			
15.	All Respiratory	.0001	15.	All Respiratory	.002
17.	Trachea-Lung	.0001	17.	Trachea-Lung	.004
			20.	Breast	.001
			24.	Ovary	.0001
25.	Male Reproductive	.003			
26.	Urinary	.001	26.	Urinary	.0001
27.	Kidney	.009	27.	Kidney	.0001
28.	Bladder	.0001	28.	Bladder	.0001
29.	Brain	.006			
			30.	Thyroid	.008
32.	Leukemia	.001	32.	Leukemia	.001
33.	Lung Small Cell CA	.002			
34.	Lung Squamous	.0001	34.	Lung Squamous	.001
35.	Lung Adenoca NOS	.0001			

SECTION 11

COMPARISON OF ORIGINAL ASBESTOS-CANCER STUDY RESULTS WITH EXPOSURE STUDY RESULTS

Table 9 shows the body sites of cancer that were significant ($p<.01$) for both the original asbestos-cancer study and the reanalysis including years of exposure. For males there is perfect agreement between the two studies with sites 13 peritoneum, 15 all respiratory, and 17 trachea-lung being significant in both cases. For females the reanalysis shows the addition of sites 2 all digestive, 5 stomach, and 20 breast to the list of significant sites and the subtraction of sites 13 peritoneum and 18 pleura from the same list. For females there was agreement between the studies of sites 9 digestive related organs, 11 gall bladder, and 12 pancreas. In both studies there is little or no agreement between the sexes in terms of cancer body sites that were found to be significantly positively related to asbestos.

A comparison between column 1 of Table 9 (original study significant sites for asbestos) and Table 8 (reanalysis study significant sites for asbestos x years of exposure) reveals that all of the sites that were significant in the original study were also significant in the reanalysis study. In the reanalysis study many more sites were found to be significant than were found in the original study. In general terms one major difference between the results of the two studies was the finding of significant urinary system cancer body sites in the reanalysis study whereas none of these sites were found to be significant in the original asbestos-cancer study. This was true for both males and females. Another major difference was the finding of significant digestive system cancer body sites for males in the reanalysis study but not in the original study. Significant respiratory system cancer body sites for females were found in the reanalysis study but not in the original study.

TYPING GUIDE SHEET

TABLE 9

SIGNIFICANT (p<.01) REGRESSION COEFFICIENTS AND SIGNIFICANCE FOR ASBESTOS FOR ORIGINAL AND REANALYSIS STUDIES BY SEX

Original Study Significant Sites (p<.01)			Reanalysis Study Significant Sites (p<.01)		
	SITE	SIGN.		SITE	SIGN.
	MALE			MALE	
13	Peritoneum	.002	13	Peritoneum	.003
15	All Respiratory	.001	15	All Respiratory	.003
17	Trachea-Lung	.0001	17	Trachea-Lung	.002
	FEMALE			FEMALE	
			2	All Digestive	.002
			5	Stomach	.002
9	Digest. Rel. Org.	.004	9	Digest. Rel. Org.	.0001
11	Gall Bladder	.001	11	Gall Bladder	.0001
12	Pancreas	.005	12	Pancreas	.001
13	Peritoneum	.009			
			20	Breast	.003

SECTION 12

DISCUSSION

The addition of the variable years of exposure to the regressions of cancers on asbestos and a number of covariables does not appear to alter the observed relationships between asbestos and cancer. These regressions do indicate that estimated exposure time is significantly positively related to cancers of almost all sites of the body. Regressions of cancer on covariables and the interaction variable asbestos x years of exposure include significant positive relationships between many cancer body sites and asbestos x years of exposure.

Interpretation of these results must be reserved until consideration is given to the possible shortcomings of estimating years of exposure by available census data. Years of exposure were estimated by the item "Year moved into unit" which was reported in the census in five time categories. Ideally the exact dates that the study population moved into their respective units would be available rather than the categorical type presented in the census. Time intervals such as those given in the categorical variable "Year moved into unit" require the application of midpoints to subjects in the intervals as the actual dates that the subjects moved into their respective units. In other words all subjects within an interval are assumed to have moved into their respective units at the midpoint of the time interval. If the subjects moved into their respective units uniformly across the time interval, then this assumption is valid from the standpoint of estimating person years from the interval to some endpoint (in this case the midpoint of the observational period for cancer incidence). The first category (1949 or earlier) of the "Year moved into unit" variable is not a closed interval however. Therefore, no midpoint exists for this category and a date within this open ended category must be chosen and used in the same manner as the midpoints of the closed intervals of the other four categories. The distribution of movers in this category is likely to be skewed such that most of the subjects moved into their units closer to 1949 than much earlier than this date. This is an assumption however that cannot be argued or supported. In the case of the present research 1/1/49 was chosen to represent the date at which subjects in this category moved into their units. Although 1/1/49 might not be the most accurate date to use in this case, any date chosen would involve guesswork and the date chosen (1/1/49) at least could be considered conservative in the sense that it would underestimate years of exposure throughout the study population if the actual date was earlier than 1/1/49.

Another possible shortcoming of the estimates of years of exposure is that these estimates reflect the age distributions of the census tracts and not just the amount of exposure time within these tracts. These shortcomings and associated assumptions point out the need for better data concerning the

moving histories of the population at the census tract level. Ideally information would be available or collected at the individual level. This is impractical in most cases and especially in studies of rare diseases that have long latent periods such as many forms of cancer such that a large number of subjects would have to be followed for long periods of time. The recommendation is that the Bureau of the Census expand its data collection to include more detailed information on moving histories at the census tract level.

If it is assumed that years of exposure as estimated here are accurate in an absolute or relative sense, then the question still remains about the environmental agent(s) to which this exposure refers. Obviously asbestos is present in this study area and the levels of asbestos to which the study population is exposed vary widely across the SMSA. Perhaps other environmental agents exist and are also related to cancer incidence. Certainly the smoking histories of the study population cannot be overlooked. This problem relates more to indirect design of the study than it does to the way in which years of exposure were measured or to the fact that years of exposure were used at all in the reanalysis.

The results of the reanalysis do corroborate the findings of the original asbestos-cancer study. When the interaction variable asbestos x years of exposure replaces the separate terms of asbestos and years of exposure in the regressions of the reanalysis, many cancer body sites are found to be significantly positively related to the interaction variable.

TYPING GUIDE SHEET

REFERENCES

1. U.S. Environmental Protection Agency, Asbestos in Domestic Water Supplies in Five California Counties. EHS Publication No. 78-2, 1979.

2. U.S. Dept. of Commerce, Bureau of the Census. 1970 Census of Population and Housing, Census Tracts San Francisco-Oakland, California SMSA. PHC(1)-189, 1972.

3. U.S. Dept. of Commerce, Bureau of the Census. 1960 Census of Population and Housing, Census Tracts San Francisco-Oakland, California SMSA. PHC(1)-137, 1962.

4. U.S. Dept. of Commerce, Bureau of the Census. 1950 Census of Population, Volume II Characteristics of Population, Part 5, California. U.S. Government Printing Office, 1952.

5. Shyrock, H. S., Siegel, J. S., et al. The Methods and Materials of Demography, Volume 2. U.S. Dept. of Commerce, 1973.

6. Chiang, C. L. Introduction to Stochastic Processes in Biostatistics. New York: Wiley, 1968.

7. Iosifescu, M. and Tautu, P. Stochastic Processes and Applications in Biology and Medicine II: Models. Berlin-Heidelberg-New York: Springer-Verlag, 1973.

8. Bureau of the Budget. Standard Metropolitan Statistical Areas. U.S. Government Printing Office, 1967.

9. Nie, N. H., et al. Statistical Package for the Social Sciences, Second Edition. New York: McGraw-Hill, 1975.